PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-108894
(43)Date of publication of application : 12.04.2002

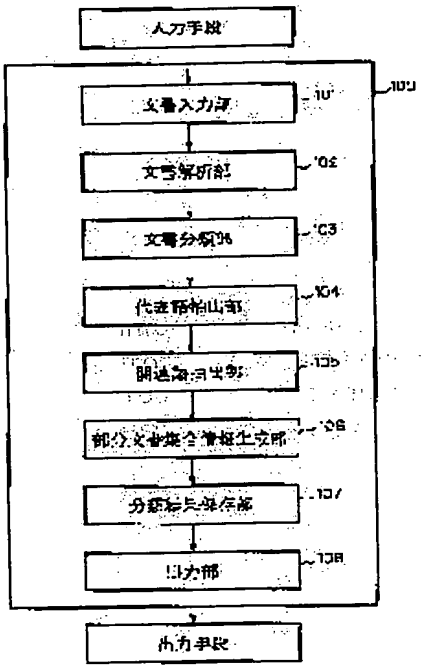
(51)Int.CI. G06F 17/30
G06F 17/21
G06F 17/28

(21)Application number : 2000-293597
(22)Date of filing : 27.09.2000

(71)Applicant : RICOH CO LTD
(72)Inventor : KENMOCHI EIJI

(54) DEVICE AND METHOD FOR SORTING DOCUMENT AND RECORDING MEDIUM FOR EXECUTING THE METHOD

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a document sorting device which can supply information effective for analyzing a partial document set and also can extract many analysis information from a document set.
SOLUTION: A document analysis part 102 extracts word information from a document that is inputted from a document input part 101 and a document sorting part 103 sorts the document into a partial document set on the basis of the word information. A keyword extraction part 104 extracts a keyword set from the partial document set and a relative word extraction part 105 extracts a relative word set of the partial document set by using a relative word dictionary. A partial document set information generation part 106 generates information on each partial document set and the relative information on these partial document sets on the basis of information on the relative word set, the keyword set and the document set of partial document set. Then a sorting result preservation part 107 preserves the sorting result of the part 103 and information generated at the part 106.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－108894
（P2002－108894A）
(43)公開日　平成14年４月12日(2002.4.12)

(51)Int.Cl.⁷		識別記号	FI		テーマコード(参考)
G06F	17/30	210	G06F	17/30	210D　5B009
		170			170A　5B075
	17/21	550		17/21	550A　5B091
	17/28			17/28	Z

審査請求　未請求　請求項の数39　ＯＬ　（全 34 頁）

(21)出願番号　特願2000－293597(P2000－293597)

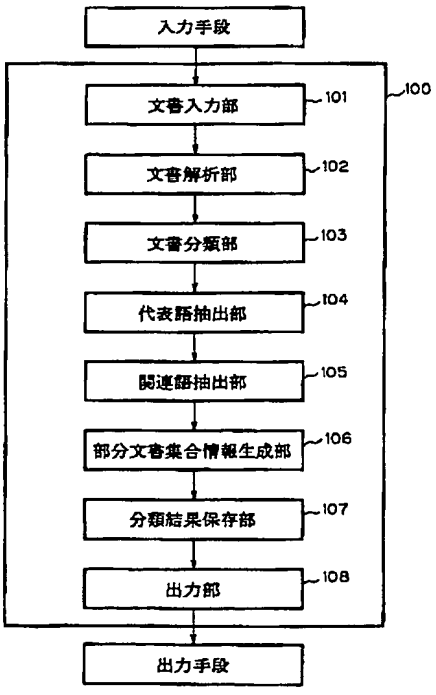
(22)出願日　平成12年９月27日(2000.9.27)

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号
(72)発明者　剣持　栄治
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(74)代理人　100079843
弁理士　高野　明近　（外２名）

Ｆターム(参考)　5B009 MC04 QA01
5B075 ND03 NK35 NR12 QP03
5B091 AB12 AB13 AB17 CA02

(54)【発明の名称】　文書分類装置、文書分類方法及び該方法を実行するための記録媒体

(57)【要約】
【課題】　部分文書集合の分析に有効な情報を提供し、文書集合からより多くの分析情報を抽出しうる文書分類装置を提供する。
【解決手段】　文書解析部１０２は、文書入力部１０１から入力した文書から単語情報を抽出し、文書分類部１０３がこれをもとに文書を部分文書集合に分類する。代表語抽出部１０４は、部分文書集合から代表語セットを抽出し、関連語抽出部１０５が、関連語辞書を用いて部分文書集合の関連語セットを抽出する。部分文書集合情報生成部１０６は、関連語セット及び代表語セットと部分文書集合の文書集合に関する情報をもとに個々の部分文書集合及びこれらの間の関連情報を生成し、分類結果保存部１０７が文書分類部１０３の分類結果と部分集合情報生成部１０６で生成された情報とをあわせて保存する。



1

【特許請求の範囲】

【請求項１】　文書集合をその内容に従って分類する文書分類装置であって、複数の文書を入力する文書入力部と、該文書入力部にて入力された各文書から該各文書を構成する単語情報を抽出する文書解析部と、該文書解析部にて抽出された各文書の単語情報をもとに前記複数の文書による文書集合をいくつかの部分文書集合に分類する文書分類部と、該文書分類部にて分類された各部分文書集合からそれらの代表語セットを抽出する代表語抽出部と、任意の単語についてその関連語が記述された関連語辞書を用いて前記代表語抽出部にて抽出した各部分文書集合の代表語セットそれぞれについて関連語セットを抽出する関連語抽出部と、該関連語抽出部にて抽出した関連語セットと前記代表語抽出部で抽出した代表語セットと各部分文書集合に所属する文書に関する情報とをもとに個々の部分文書集合及び部分文書集合間の関連情報を生成する部分文書集合情報生成部と、前記文書分類部での分類結果を前記部分文書集合情報生成部にて生成された情報と合わせて保存する分類結果保存部とを含むことを特徴とする文書分類装置。

【請求項２】　請求項１に記載の文書分類装置において、前記関連語抽出部にて抽出される関連語セットが、同義語、類義語、反対語のうちの少なくとも一つ以上の組合わせであることを特徴とする文書分類装置。

【請求項３】　請求項１に記載の文書分類装置において、前記関連語抽出部にて抽出される関連語セットが少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成部をさらに含むことを特徴とする文書分類装置。

【請求項４】　請求項１に記載の文書分類装置において、前記関連語抽出部にて抽出される関連語が少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットと代表語セットから反対語セットに対応する代表語を除いた単語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成部をさらに含むことを特徴とする文書分類装置。

【請求項５】　文書の内容に従って文書の分類を行う文書分類装置であって、文書データを入力する文書入力部と、前記文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析部と、該文書解析部にて抽出された文書データの解析情報から文書データを多次元ベクトル空間で表現するための文書ベクトル空間を生成する文書ベクトル空間生成部と、該文書ベクトル空間生成部にて生成した文書ベクトル空間において統計手法を用いることにより文書データの分類を行う文書分類部とを含み、前記文書解析部にて抽出される特定の品詞を有する単語を、該特定の品詞の品詞情報に基づき、該特定の品詞の前後に抽出される一つ以上の単語と結合することにより生成される単語と置き換え、かつ該特定の品詞の品詞情報も適切に置き換えることを特徴とする文書分類装置。

2

【請求項６】　請求項５に記載の文書分類装置において、前記文書分類部において統計手法としてクラスタリング法を用いることで文書データの分類を行うことを特徴とする文書分類装置。

【請求項７】　請求項５または６に記載の文書分類装置において、前記文書解析部において品詞が接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、単語および品詞の置き換えを行うことを特徴とする文書分類装置。

【請求項８】　請求項５ないし７のいずれか１に記載の文書分類装置において、前記文書解析部において特定の品詞の単語が出現するまで単語の結合を続けることを特徴とする文書分類装置。

【請求項９】　請求項５ないし８のいずれか１に記載の文書分類装置において、前記文書解析部において品詞が数詞接尾詞もしくは助数詞の単語について、該数詞接尾詞もしくは助数詞の単語に結合される複数の単語を削除し、前記文書分類部では削除した単語の情報を用いないことを特徴とする文書分類装置。

【請求項１０】　文書の内容に従って文書データ集合を分類する文書分類装置であって、文書データ集合を入力する文書入力部と、すべての文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析部と、該文書解析部にて抽出された文書データの解析結果を記憶する文書解析結果記憶部と、前記文書解析部にて抽出された文書データの解析情報から前記文書データを多次元ベクトル空間で表現するためのベクトル空間を生成する文書ベクトル空間生成部と、該文書ベクトル空間生成部にて生成された文書ベクトル空間の各文書データのベクトルデータを記憶する文書ベクトルデータ記憶部と、指定される条件から文書データ集合の分類数を決定する分類数決定部と、前記文書ベクトル空間生成部にて生成した文書ベクトル空間において統計手法を用いることにより文書データを前記指定された分類数の部分文書集合に分類する文書分類部と、該文書分類部で生成された分類結果を記憶する分類結果記憶部と、前記分類数決定部から前記分類結果記憶部までの処理を繰り返し行うか否かの判定をおこなう繰り返し判定部と、前記文書ベクトルデータ記憶部と前記分類結果記憶部に記憶された情報を用いて生成されたすべての部分文書集合間の関係情報を算出する部分文書集合間関係算出部と、該部分文書集合間関係算出

部にて生成された部分文書集合間の関係情報を記憶する部分文書集合間関係記憶部とを含むことを特徴とする文書分類装置。

【請求項１１】　請求項１０に記載の文書分類装置において、前記文書分類部にて用いられる統計手法が非階層クラスタリング手法であることを特徴とする文書分類装置。

【請求項１２】　請求項１０または１１に記載の文書分類装置において、前記部分文書集合間関係算出部にて算出される関係が、類似関係と包含関係であることを特徴とする文書分類装置。

【請求項１３】　請求項１２に記載の文書分類装置において、前記部分文書集合間の関係は各部分文書集合から抽出される単語情報のみを用いて算出されることを特徴とする文書分類装置。

【請求項１４】　文書集合をその内容に従って分類する文書分類方法であって、複数の文書を入力する文書入力ステップと、該文書入力ステップにて入力された各文書から該各文書を構成する単語情報を抽出する文書解析ステップと、該文書解析ステップにて抽出された各文書の単語情報をもとに前記複数の文書による文書集合をいくつかの部分文書集合に分類する文書分類ステップと、該文書分類ステップにて分類された各部分文書集合からそれらの代表語セットを抽出する代表語抽出ステップと、任意の単語についてその関連語が記述された関連語辞書を用いて前記代表語抽出ステップにて抽出した各部分文書集合の代表語セットそれぞれについて関連語セットを抽出する関連語抽出ステップと、該関連語抽出ステップにて抽出した関連語セットと前記代表語抽出ステップで抽出した代表語セットと各部分文書集合に所属する文書に関する情報とをもとに個々の部分文書集合及び部分文書集合間の関連情報を生成する部分文書集合情報生成ステップと、前記文書分類ステップでの分類結果を前記部分文書集合情報生成部にて生成された情報と合わせて保存する分類結果保存ステップとを含むことを特徴とする文書分類方法。

【請求項１５】　請求項１４に記載の文書分類方法において、前記関連語抽出ステップにて抽出される関連語セットが、同義語、類義語、反対語のうちの少なくとも一つ以上の組合わせであることを特徴とする文書分類方法。

【請求項１６】　請求項１４に記載の文書分類方法において、前記関連語抽出ステップにて抽出される関連語セットが少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含むことを特徴とする文書分類方法。

【請求項１７】　請求項１４に記載の文書分類方法において、前記関連語抽出ステップにて抽出される関連語が少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットと代表語セットから反対語セットに対応する代表語を除いた単語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含むことを特徴とする文書分類方法。

【請求項１８】　文書の内容に従って文書の分類を行う文書分類方法であって、文書データを入力する文書入力ステップと、前記文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析情報から文書データを多次元ベクトル空間で表現するための文書ベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データの分類を行う文書分類ステップとを含み、前記文書解析ステップにて抽出される特定の品詞を有する単語を、該特定の品詞の品詞情報に基づき、該特定の品詞の前後に抽出される一つ以上の単語と結合することにより生成される単語と置き換え、かつ該特定の品詞の品詞情報も適切に置き換えることを特徴とする文書分類装置。

【請求項１９】　請求項１８に記載の文書分類方法において、前記文書分類ステップにおいて統計手法としてクラスタリング法を用いることで文書データの分類を行うことを特徴とする文書分類方法。

【請求項２０】　請求項１８または１９に記載の文書分類方法において、前記文書解析ステップにおいて品詞が接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、単語及び品詞の置き換えを行うことを特徴とする文書分類方法。

【請求項２１】　請求項１８ないし２０のいずれか１に記載の文書分類方法において、前記文書解析ステップにおいて特定の品詞の単語が出現するまで単語の結合を続けることを特徴とする文書分類方法。

【請求項２２】　請求項１８ないし２１のいずれか１に記載の文書分類方法において、前記文書解析ステップにおいて品詞が数詞接尾詞もしくは助数詞の単語について、該数詞接尾詞もしくは助数詞の単語に結合される複数の単語を削除し、前記文書分類ステップでは削除した単語の情報を用いないことを特徴とする文書分類方法。

【請求項２３】　文書の内容に従って文書データ集合を分類する文書分類方法であって、文書データ集合を入力する文書入力ステップと、すべての文書データに形態素

解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析結果を記憶する文書解析結果記憶ステップと、前記文書解析ステップにて抽出された文書データの解析情報から前記文書データを多次元ベクトル空間で表現するためのベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成された文書ベクトル空間の各文書データのベクトルデータを記憶する文書ベクトルデータ記憶ステップと、指定される条件から文書データ集合の分類数を決定する分類数決定ステップと、前記文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データを前記指定された分類数の部分文書集合に分類する文書分類ステップと、該文書分類ステップで生成された分類結果を記憶する分類結果記憶ステップと、前記分類数決定ステップから前記分類結果記憶ステップまでの処理を繰り返し行うか否かの判定をおこなう繰り返し判定ステップと、前記文書ベクトルデータ記憶ステップと前記分類結果記憶ステップにて記憶された情報を用いて生成されたすべての部分文書集合間の関係情報を算出する部分文書集合間関係算出ステップと、該部分文書集合間関係算出ステップにて生成された部分文書集合間の関係情報を記憶する部分文書集合間関係記憶ステップとを含むことを特徴とする文書分類方法。

【請求項２４】　請求項２３に記載の文書分類方法において、前記文書分類ステップにて用いられる統計手法が非階層クラスタリング手法であることを特徴とする文書分類方法。

【請求項２５】　請求項２３または２４に記載の文書分類方法において、前記部分文書集合間関係算出ステップにて算出される関係が、類似関係と包含関係であることを特徴とする文書分類方法。

【請求項２６】　請求項２５に記載の文書分類方法において、前記部分文書集合間の関係は各部分文書集合から抽出される単語情報のみを用いて算出されることを特徴とする文書分類方法。

【請求項２７】　文書集合をその内容に従って分類する文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、複数の文書を入力する文書入力ステップと、該文書入力ステップにて入力された各文書から該各文書を構成する単語情報を抽出する文書解析ステップと、該文書解析ステップにて抽出された各文書の単語情報をもとに前記複数の文書による文書集合をいくつかの部分文書集合に分類する文書分類ステップと、該文書分類ステップにて分類された各部分文書集合からそれらの代表語セットを抽出する代表語抽出ステップと、任意の単語についてその関連語が記述された関連語辞書を用いて前記代表語抽出ステップにて抽出した各部分文書集合の代表語セットそれぞれについて関連語セットを抽出する関連語抽出ステップと、該関連語抽出ステップにて抽出した関連語セットと前記代表語抽出ステップで抽出した代表語セットと各部分文書集合に所属する文書に関する情報とをもとに個々の部分文書集合及び部分文書集合間の関連情報を生成する部分文書集合情報生成ステップと、前記文書分類ステップでの分類結果を前記部分文書集合情報生成部にて生成された情報と合わせて保存する分類結果保存ステップとを含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項２８】　請求項２７に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記関連語抽出ステップにて抽出される関連語セットが、同義語、類義語、反対語のうちの少なくとも一つ以上の組合わせである文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項２９】　請求項２７に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記関連語抽出ステップにて抽出される関連語セットが少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項３０】　請求項２７に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記関連語抽出ステップにて抽出される関連語が少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットと代表語セットから反対語セットに対応する代表語を除いた単語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項３１】　文書の内容に従って文書の分類を行う文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、文書データを入力する文書入力ステップと、前記文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析情報から文書データを多次元ベクトル空間で表現する

ための文書ベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データの分類を行う文書分類ステップとを含み、前記文書解析ステップにて抽出される特定の品詞を有する単語を、該特定の品詞の品詞情報に基づき、該特定の品詞の前後に抽出される一つ以上の単語と結合することにより生成される単語と置き換え、かつ該特定の品詞の品詞情報も適切に置き換える文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項３２】　請求項３１に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書分類ステップにおいて統計手法としてクラスタリング法を用いることで文書データの分類を行う文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項３３】　請求項３１または３２に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書解析ステップにおいて品詞が接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、単語及び品詞の置き換えを行う文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項３４】　請求項３１ないし３３のいずれか１に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書解析ステップにおいて特定の品詞の単語が出現するまで単語の結合を続ける文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項３５】　請求項３１ないし３４のいずれか１に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書解析ステップにおいて品詞が数詞接尾詞もしくは助数詞の単語について、該数詞接尾詞もしくは助数詞の単語に結合される複数の単語を削除し、前記文書分類ステップでは削除した単語の情報を用いない文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項３６】　文書の内容に従って文書データ集合を分類する文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、文書データ集合を入力する文書入力ステップと、すべての文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析結果を記憶する文書解析結果記憶ステップと、前記文書解析ステップにて抽出された文書データの解析情報から前記文書データを多次元ベクトル空間で表現するためのベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成された文書ベクトル空間の各文書データのベクトルデータを記憶する文書ベクトルデータ記憶ステップと、指定される条件から文書データ集合の分類数を決定する分類数決定ステップと、前記文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データを前記指定された分類数の部分文書集合に分類する文書分類ステップと、該文書分類ステップで生成された分類結果を記憶する分類結果記憶ステップと、前記分類数決定ステップから前記分類結果記憶ステップまでの処理を繰り返し行うか否かの判定をおこなう繰り返し判定ステップと、前記文書ベクトルデータ記憶ステップと前記分類結果記憶ステップにて記憶された情報を用いて生成されたすべての部分文書集合間の関係情報を算出する部分文書集合間関係算出ステップと、該部分文書集合間関係算出ステップにて生成された部分文書集合間の関係情報を記憶する部分文書集合間関係記憶ステップとを含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項３７】　請求項３６に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書分類ステップにて用いられる統計手法が非階層クラスタリング手法である文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項３８】　請求項３６または３７に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記部分文書集合間関係算出ステップにて算出される関係が、類似関係と包含関係である文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項３９】　請求項３８に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記部分文書集合間の関係は各部分文書集合から抽出される単語情報のみを用いて算出される文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、文書分類装置、文書分類方法及び該方法を実行するための記録媒体に関し、情報分類、情報分析、情報検索等に応用可能な文書分類技術に関する。

【０００２】

【従来の技術】インターネット等の普及により大量の文書情報へのアクセスが可能になり、収集した大量の文書情報を意味のあるグループに分類し、文書集合の構造を

把握するなどの知的作業が行われ始めている。大量な文書集合を分析する場合、まず文書集合をいくつかの話題で分類し、得られた部分文書集合（ある基準で集められた複数の文書）を単位としてさまざまな作業を行うことで、分析作業を効率的に行うことができるものと考えられる。大量の文書情報をユーザが手動で分類する場合、人的／時間的コストが膨大なものになるため、文書集合を文書の内容により自動分類できる装置が望まれている。

【０００３】従来、膨大な文書集合からの質の高い分類結果を得るための発明が広く行われている。例えば、特開平７－３６８９７号公報に記載の発明は、分類対象文書集合に含まれる単語を特徴量とする文書特徴ベクトルを用い、その文書特徴ベクトルに対してクラスタリング手法を適用して分類を行うものである。上記の発明ではユーザの意図を反映した分類を行うためにクラスタリングの初期重心ベクトルをユーザが指定することも示唆している。

【０００４】また、特開平１１－２９６５５２号公報に記載の発明は、単語の多義性/同義性を考慮するために文書間の内積行列に特異値分解を適用することにより文書間の単語の共起性を基に潜在的意味空間を生成して、文書と単語を潜在的意味空間に射影し、その潜在的意味空間においてクラスタリング手法などを用いて文書分類を行うものである。このように膨大な文書集合からの質の高い分類結果を得るための発明は種々提案されているが、文書集合の分析を行うためには文書集合を分類するだけは不十分であり、生成された部分文書集合からどのように有効な情報を抽出するかということも重要な問題であるが、この点についての発明はあまり見られない。

【０００５】また、形態素解析などの自然言語処理を用いて文書からそれらを構成する単語を抽出することにより文書を単語頻度のベクトル（文書特徴ベクトル）として空間表現することが可能となるが、これは文書ベクトル空間モデルと呼ばれ、広く用いられている。上述した特開平７－３６８９７号公報の発明は、このような文書ベクトル空間において、クラスタリング手法を適用することにより文書分類を行うものである。

【０００６】このように文書ベクトル空間で統計的手法を用いて文書分類処理や文書検索処理等を行う場合、文書ベクトル空間が異なれば得られる結果の質も変わると考えられるので、如何にして良い文書ベクトル空間を生成するかが高品位な処理結果を得るためには重要な問題となる。

【０００７】前述したように、通常文書ベクトル空間の各軸は分類対象文書データに形態素解析を適用した結果抽出される単語をもとに構成されるため、例えば、特開平１１－１１０４０８号公報や特開平１１－２５９４８７号公報に代表される発明は、検索問い合わせ語や検索対象文書に対し、形態素解析を適用し、その結果抽出される単語から適切な条件のもとに複合語を生成し、これらの複合語の情報も前記文書ベクトル空間の生成に用いることで、文書ベクトル空間上で行う文書検索の精度の向上を目的としている。従って、文書ベクトル空間で文書分類処理を行う場合においても、複合語を考慮して文書ベクトル空間を生成することで高品位な分類結果を得ることが期待される。

【０００８】ところで、上記先願を含め、通常複合語を考慮する場合は、品詞が名詞もしくはそれに類するものが対象とされているが、名詞だけでなく他の結合可能な品詞も適切に結合させることで、より高品位な文書ベクトル空間を構成することが可能になると考えられる。すなわち、先願の発明等ではあまり扱われることのなかった、接頭詞、接尾詞、助数詞、及びそれらに類する品詞を有する単語について、適切な基準でそれらの前後の単語と結合することで生成される単語と置き換えるとともに、品詞も適切なものに置き換えることを考える。

【０００９】例えば、“イタリア製の車”という文字列に対して形態素解析を適用し、“イタリア［普通名詞］、製［接頭詞］、の［格助詞］、車”という結果が得られた場合、接頭詞である“製”という単語に着目し、これをこの直前に抽出されている“イタリア”という普通名詞と結合し、“イタリア製”という単語を生成し、これを普通名詞の品詞を有する単語として、“製”という単語と置き換える。そして、この文字列に加え、“イタリアの特色”、“イタリア製の皿”という文字列で構成するベクトル空間を生成することを考えてみる。

【００１０】名詞だけで空間を生成することを考えた場合、前記の結合・置き換え処理を行わない場合、ベクトル空間を構成する単語は、“イタリア、車、特色、皿”であり、前記文字列は、単語の出現頻度を座標値と考えた場合、（１，０，１，０）、（１，０，０，０）、（１，０，０，１）となる。この場合、前記３つの文字列の相互の類似度をベクトル間の内積で計算すると、前記３つの文字列の相互の類似度は同じものとなる。一方、前記の結合・置き換え処理を行った後、名詞で空間を生成すると、ベクトル空間を構成する単語は、“イタリア、イタリア製、車、特色、皿”となる。同様に、前記３文書のこの空間でのベクトルは、（１，１，０，１，０）、（１，０，０，０，０）、（１，１，０，０，１）となる。この場合、前記３つの文字列の相互の類似度には、差異が生じ、最初の文字列と最後の文字列が２番目の文字列より高い類似度を持つことになる。すなわち、この結合・置き換え処理によりベクトル空間により限定化された意味を測る特徴次元を加えることができ、これによりこのベクトル空間で行う文書分類等の質も向上するものと考えられる。

【００１１】また、“２０００年の目標”という文字列に対し形態素解析を適用し、“２０００［数詞］、年［助数詞］、の［格助詞］、目標［普通名詞］”のよう

な結果が得られているとする。このとき、助数詞である“年”という単語に着目し、これをこの直前に抽出されている“２０００”という数詞と結合し、“２０００年”という単語を生成し、これを普通名詞の品詞を有する単語として、“年”という単語と置き換え、かつ“２０００”という数詞を削除する。これにより、非常に漠然とした意味しか有していない助数詞である“年”や“２０００”という単語にかえて、“２０００年”というより意味的に限定された、それゆえ変数としてはより重要な単語をもとにして文書ベクトル空間が構成可能になることが期待される。

【００１２】また、上述のようにインターネット等の普及により大量の文書データへのアクセスが可能になり、その結果として興味のある情報が記述されている文書データを簡単にかつ大量に収集できるようになったが、しかしその一方で、収集した文書データが大量であるがために、それら文書データから有効な情報を読み取る作業は非常に困難なものになってしまっている。このため、大量の文書データから自動もしくは半自動で有効な情報を簡単に抽出することを目的として、文書検索や文書自動分類に関する研究・開発が盛んに行われている。特に、文書分類手法は、生成される複数の部分文書データ集合個々を文書データに含まれる複数の話題を示すものと考えると、文書データ全体の構造を把握する手法として非常に有効なものである。

【００１３】上述のような目的のために開発された手法の代表的なものに、Scatter/Gather法（D.Cutting et.al., Scater/Gather: A Cluster-based Approach to Browsing Large Document Collections., Proc. ACM SIGIR ’92）がある。Scatter/Gather法では、文書データ集合の話題を代表文書と代表単語により表現するとともに、話題が不明瞭な文書集合に対して逐次クラスタリングを適用し、複数の部分文書データ集合に分割していくことで文書集合に含まれる様々な話題を理解していく。文書集合の構造を理解するためには、文書集合に含まれる部分文書集合個々を理解することはもちろん必要であるが、加えて部分文書集合間の関係に関する情報も必要であると考えられる。しかしながら、Scatter/Gather法では個々の部分文書集合に関する情報しか提示されていないため、Scatter/Gather法のみでは文書集合の構造を把握することは困難であると考えられる。

【００１４】また、一般的に文書分類手法においては、生成する部分文書集合の数が実行前に必要であるが、最適な部分文書集合の数を予測することは極めて困難である。しかも、一方で生成する部分文書集合の数が異なれば生成される部分文書集合の構造も変化してしまう。このため、必要な情報を得るためには生成する部分文書集合の数をかえながら、繰り返し文書分類を行わなければならない。Scatter/Gather法はこの点についても一つの解決法を提示しており、ユーザがより詳細な構造を知

りたいと考える部分文書集合のみに対し逐次クラスタリングを適用し、あらたな部分文書集合を生成し、それらを詳細に分析することで所望の情報を得ることができるとともに、この行為により文書集合全体の構造を理解することも容易になっていると考えられる。

【００１５】すなわち、ユーザが行いたいことは文書集合の構造の把握であり、部分文書集合を生成するという行為は本来ユーザが行う必要がないものと考えられる。そして、ユーザが、事前に文書集合から様々な数の部分文書集合を生成し、生成された多数の部分文書集合間の関係を算出しておくことで、ユーザは始めから構造の把握を行う作業に集中できると考えられる。しかしながら、前述の通りScatter/Gather法は部分文書間の関連に関しては考慮されていない。

【００１６】

【発明が解決しようとする課題】本発明の請求項１～４，１４～１７，及び２７～３０の発明では、文書分類を行うとともに、単語の関連語情報を基に生成された部分文書集合個々及びそれらの関連情報をさらに生成することで部分文書集合の分析に有効な情報を提供することを目的とする。さらに、関連語として反対語に着目することで、各部分文書集合の代表語セットの反対語を含む部分文書集合が、生成された部分文書集合にはない場合、反対語を含むあらたな部分文書集合を生成することで、文書集合からより多くの分析情報を抽出しうる文書分類装置を提供することを目的とする。

【００１７】従って請求項１，１４及び２７の発明は、生成された部分文書集合それぞれの代表語セットを抽出し、さらにそれら代表語それぞれについて関連語を求め、これらの情報をもとに各部分文書集合および部分文書集合間の関連情報を生成することで、部分文書集合の分析に有効な情報を提供する文書分類装置、方法または記録媒体を提供することを目的とする。

【００１８】請求項２，１５及び２８の発明は、関連語として同義語、類義語、反対語のすくなくとも一つ以上の組合わせを用いることで主に類似性に関す情報を提供する文書分類装置、方法または記録媒体を提供することを目的とする。

【００１９】請求項３，４，１６，１７，２９及び３０の発明は、各部分文書集合の代表語セットの関連語として反対語を用い、反対語が自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、その反対語を含む文書を文書集合から抽出し、それを新たな部分文書集合とすることで、文書集合からより多くの分析情報を抽出する文書分類装置、方法または記録媒体を提供することを目的とする。

【００２０】また請求項５，１８及び３１の発明は、分類対象文書に形態素解析し、得られた解析結果をもとに分類対象文書を幾つかの文書集合に分類する文書分類装置において、形態素解析の結果得られる単語のうち指定

される品詞をもつ単語について、その前後の単語と適切に組合わせた単語と置き換え、かつ品詞もまた適切なものに置き換える処理を施すことによって、高品位な文書ベクトル空間を構成し、この文書ベクトル空間で統計処理を用いて文書分類を行うことで高品質な文書分類結果を得ることができる文書分類装置、方法または記録媒体を提供することを目的とする。

【００２１】請求項６，１９及び３２の発明は、文書分類を行うための統計手法として、クラスタリング手法を用いることで、簡便に高品質な文書分類結果を得ることができる文書分類装置、方法または記録媒体を提供することを目的とする。

【００２２】請求項７，２０及び３３の発明は、分類対象文書に形態素解析を適用することで抽出される単語の中で、特に、品詞が、接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、適切な結合処理を施こすことで、高品質な文書ベクトル空間を得ることができる文書分類装置、方法または記録媒体を提供することを目的とする。

【００２３】請求項８，２１及び３４の発明は、単語の結合処理において特定の品詞の単語が出現するまで単語の結合を続けることによって新たな単語を生成することで、高品質な文書ベクトル空間を得ることができる文書分類装置、方法または記録媒体を提供することを目的とする。

【００２４】請求項９，２２及び３５の発明は、単語の結合処理において、品詞が数詞接尾詞もしくは助数詞の単語について、結合される複数の単語を削除し、文書ベクトル空間を生成する際にはそれらの単語の情報は用いないことで、高品質な文書ベクトル空間を得ることができる文書分類装置、方法または記録媒体を提供することを目的とする。

【００２５】また本発明の請求項１０～１３，２３～２６及び３６～３９の発明では、事前に文書集合から様々な数の部分文書集合を生成し、生成された多数の部分文書集合間の関係を算出することで、ユーザが始めから文書集合の構造の把握を行う作業に集中できる情報を提供することを目的とする。

【００２６】従って請求項１０，２３及び３６の発明は、文書のベクトル空間モデルを用い、生成する部分文書集合の数をパラメータとして繰り返し文書分類処理を行うことで、多数の部分文書集合を生成し、さらに生成された多数の文書集合について相互の関係を算出することで、文書集合の構造の把握を支援しうる情報を生成する文書分類装置を提供する文書分類装置、方法または記録媒体を提供することを目的とする。

【００２７】請求項１１，２４及び３７の発明は、文書分類を行う統計手法として、非階層クラスタリング手法を用いることで、簡便に多数の部分文書集合を生成する文書分類装置、方法または記録媒体を提供することを目的とする。

【００２８】請求項１２，２５及び３８の発明は、生成された多数の文書集合について相互の関係として、類似関係と包含関係を算出することで、容易に文書集合の構造を把握しうる情報を提供する文書分類装置、方法または記録媒体を提供することを目的とする。

【００２９】請求項１３，２６及び３９の発明は、生成された多数の文書集合が有する情報のうち、単語に関する情報のみを用いて相互の関係を算出することで、汎用性・再利用性の高い関係情報を算出する文書分類装置、方法または記録媒体を提供することを目的とする。

【００３０】

【課題を解決するための手段】請求項１の発明は、文書集合をその内容に従って分類する文書分類装置であって、複数の文書を入力する文書入力部と、該文書入力部にて入力された各文書から該各文書を構成する単語情報を抽出する文書解析部と、該文書解析部にて抽出された各文書の単語情報をもとに前記複数の文書による文書集合をいくつかの部分文書集合に分類する文書分類部と、該文書分類部にて分類された各部分文書集合からそれらの代表語セットを抽出する代表語抽出部と、任意の単語についてその関連語が記述された関連語辞書を用いて前記代表語抽出部にて抽出した各部分文書集合の代表語セットそれぞれについて関連語セットを抽出する関連語抽出部と、該関連語抽出部にて抽出した関連語セットと前記代表語抽出部で抽出した代表語セットと各部分文書集合に所属する文書に関する情報とをもとに個々の部分文書集合及び部分文書集合間の関連情報を生成する部分文書集合情報生成部と、前記文書分類部での分類結果を前記部分文書集合情報生成部にて生成された情報と合わせて保存する分類結果保存部とを含むことを特徴としたものである。

【００３１】請求項２の発明は、請求項１の発明において、前記関連語抽出部にて抽出される関連語セットが、同義語、類義語、反対語のうちの少なくとも一つ以上の組合わせであることを特徴としたものである。

【００３２】請求項３の発明は、請求項１の発明において、前記関連語抽出部にて抽出される関連語セットが少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成部をさらに含むことを特徴としたものである。

【００３３】請求項４の発明は、請求項１の発明において、前記関連語抽出部にて抽出される関連語が少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しな

い反対語セットと代表語セットから反対語セットに対応する代表語を除いた単語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成部をさらに含むことを特徴としたものである。

【００３４】請求項５の発明は、文書の内容に従って文書の分類を行う文書分類装置であって、文書データを入力する文書入力部と、前記文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析部と、該文書解析部にて抽出された文書データの解析情報から文書データを多次元ベクトル空間で表現するための文書ベクトル空間を生成する文書ベクトル空間生成部と、該文書ベクトル空間生成部にて生成した文書ベクトル空間において統計手法を用いることにより文書データの分類を行う文書分類部とを含み、前記文書解析部にて抽出される特定の品詞を有する単語を、該特定の品詞の品詞情報に基づき、該特定の品詞の前後に抽出される一つ以上の単語と結合することにより生成される単語と置き換え、かつ該特定の品詞の品詞情報も適切に置き換えることを特徴としたものである。

【００３５】請求項６の発明は、請求項５の発明において、前記文書分類部において統計手法としてクラスタリング法を用いることで文書データの分類を行うことを特徴としたものである。

【００３６】請求項７の発明は、請求項５または６の発明において、前記文書解析部において品詞が接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、単語および品詞の置き換えを行うことを特徴としたものである。

【００３７】請求項８の発明は、請求項５ないし７のいずれか１の発明において、前記文書解析部において特定の品詞の単語が出現するまで単語の結合を続けることを特徴としたものである。

【００３８】請求項９の発明は、請求項５ないし８のいずれか１の発明において、前記文書解析部において品詞が数詞接尾詞もしくは助数詞の単語について、該数詞接尾詞もしくは助数詞の単語に結合される複数の単語を削除し、前記文書分類部では削除した単語の情報を用いないことを特徴としたものである。

【００３９】請求項１０の発明は、文書の内容に従って文書データ集合を分類する文書分類装置であって、文書データ集合を入力する文書入力部と、すべての文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析部と、該文書解析部にて抽出された文書データの解析結果を記憶する文書解析結果記憶部と、前記文書解析部にて抽出された文書データの解析情報から前記文書データを多次元ベクトル空間で表現するためのベクトル空間を生成する文書ベクトル空間生成部と、該文書ベクトル空間

生成部にて生成された文書ベクトル空間の各文書データのベクトルデータを記憶する文書ベクトルデータ記憶部と、指定される条件から文書データ集合の分類数を決定する分類数決定部と、前記文書ベクトル空間生成部にて生成した文書ベクトル空間において統計手法を用いることにより文書データを前記指定された分類数の部分文書集合に分類する文書分類部と、該文書分類部で生成された分類結果を記憶する分類結果記憶部と、前記分類数決定部から前記分類結果記憶部までの処理を繰り返し行うか否かの判定をおこなう繰り返し判定部と、前記文書ベクトルデータ記憶部と前記分類結果記憶部に記憶された情報を用いて生成されたすべての部分文書集合間の関係情報を算出する部分文書集合間関係算出部と、該部分文書集合間関係算出部にて生成された部分文書集合間の関係情報を記憶する部分文書集合間関係記憶部とを含むことを特徴としたものである。

【００４０】請求項１１の発明は、請求項１０の発明において、前記文書分類部にて用いられる統計手法が非階層クラスタリング手法であることを特徴としたものである。

【００４１】請求項１２の発明は、請求項１０または１１の発明において、前記部分文書集合間関係算出部にて算出される関係が、類似関係と包含関係であることを特徴としたものである。

【００４２】請求項１３の発明は、請求項１２の発明において、前記部分文書集合間の関係は各部分文書集合から抽出される単語情報のみを用いて算出されることを特徴としたものである。

【００４３】請求項１４の発明は、文書集合をその内容に従って分類する文書分類方法であって、複数の文書を入力する文書入力ステップと、該文書入力ステップにて入力された各文書から該各文書を構成する単語情報を抽出する文書解析ステップと、該文書解析ステップにて抽出された各文書の単語情報をもとに前記複数の文書による文書集合をいくつかの部分文書集合に分類する文書分類ステップと、該文書分類ステップにて分類された各部分文書集合からそれらの代表語セットを抽出する代表語抽出ステップと、任意の単語についてその関連語が記述された関連語辞書を用いて前記代表語抽出ステップにて抽出した各部分文書集合の代表語セットそれぞれについて関連語セットを抽出する関連語抽出ステップと、該関連語抽出ステップにて抽出した関連語セットと前記代表語抽出ステップで抽出した代表語セットと各部分文書集合に所属する文書に関する情報とをもとに個々の部分文書集合及び部分文書集合間の関連情報を生成する部分文書集合情報生成ステップと、前記文書分類ステップでの分類結果を前記部分文書集合情報生成部にて生成された情報と合わせて保存する分類結果保存ステップとを含むことを特徴としたものである。

【００４４】請求項１５の発明は、請求項１４の発明に

おいて、前記関連語抽出ステップにて抽出される関連語セットが、同義語、類義語、反対語のうちの少なくとも一つ以上の組合わせであることを特徴としたものである。

【００４５】請求項１６の発明は、請求項１４の発明において、前記関連語抽出ステップにて抽出される関連語セットが少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含むことを特徴としたものである。

【００４６】請求項１７の発明は、請求項１４の発明において、前記関連語抽出ステップにて抽出される関連語が少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットと代表語セットから反対語セットに対応する代表語を除いた単語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含むことを特徴としたものである。

【００４７】請求項１８の発明は、文書の内容に従って文書の分類を行う文書分類方法であって、文書データを入力する文書入力ステップと、前記文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析情報から文書データを多次元ベクトル空間で表現するための文書ベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データの分類を行う文書分類ステップとを含み、前記文書解析ステップにて抽出される特定の品詞を有する単語を、該特定の品詞の品詞情報に基づき、該特定の品詞の前後に抽出される一つ以上の単語と結合することにより生成される単語と置き換え、かつ該特定の品詞の品詞情報も適切に置き換えることを特徴としたものである。

【００４８】請求項１９の発明は、請求項１８の発明において、前記文書分類ステップにおいて統計手法としてクラスタリング法を用いることで文書データの分類を行うことを特徴としたものである。

【００４９】請求項２０の発明は、請求項１８または１９の発明において、前記文書解析ステップにおいて品詞が接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、単語および品詞の置き換えを行うことを特徴としたものである。

【００５０】請求項２１の発明は、請求項１８ないし２０のいずれか１の発明において、前記文書解析ステップにおいて特定の品詞の単語が出現するまで単語の結合を続けることを特徴としたものである。

【００５１】請求項２２の発明は、請求項１８ないし２１のいずれか１の発明において、前記文書解析ステップにおいて品詞が数詞接尾詞もしくは助数詞の単語について、該数詞接尾詞もしくは助数詞の単語に結合される複数の単語を削除し、前記文書分類ステップでは削除した単語の情報を用いないことを特徴としたものである。

【００５２】請求項２３の発明は、文書の内容に従って文書データ集合を分類する文書分類方法であって、文書データ集合を入力する文書入力ステップと、すべての文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析結果を記憶する文書解析結果記憶ステップと、前記文書解析ステップにて抽出された文書データの解析情報から前記文書データを多次元ベクトル空間で表現するためのベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成された文書ベクトル空間の各文書データのベクトルデータを記憶する文書ベクトルデータ記憶ステップと、指定される条件から文書データ集合の分類数を決定する分類数決定ステップと、前記文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データを前記指定された分類数の部分文書集合に分類する文書分類ステップと、該文書分類ステップで生成された分類結果を記憶する分類結果記憶ステップと、前記分類数決定ステップから前記分類結果記憶ステップまでの処理を繰り返し行うか否かの判定をおこなう繰り返し判定ステップと、前記文書ベクトルデータ記憶ステップと前記分類結果記憶ステップにて記憶された情報を用いて生成されたすべての部分文書集合間の関係情報を算出する部分文書集合間関係算出ステップと、該部分文書集合間関係算出ステップにて生成された部分文書集合間の関係情報を記憶する部分文書集合間関係記憶ステップとを含むことを特徴としたものである。

【００５３】請求項２４の発明は、請求項２３の発明において、前記文書分類ステップにて用いられる統計手法が非階層クラスタリング手法であることを特徴としたものである。

【００５４】請求項２５の発明は、請求項２３または２４の発明において、前記部分文書集合間関係算出ステップにて算出される関係が、類似関係と包含関係であることを特徴としたものである。

【００５５】請求項２６の発明は、請求項２５の発明において、前記部分文書集合間の関係は各部分文書集合か

ら抽出される単語情報のみを用いて算出されることを特徴としたものである。

【００５６】請求項２７の発明は、文書集合をその内容に従って分類する文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、複数の文書を入力する文書入力ステップと、該文書入力ステップにて入力された各文書から該各文書を構成する単語情報を抽出する文書解析ステップと、該文書解析ステップにて抽出された各文書の単語情報をもとに前記複数の文書による文書集合をいくつかの部分文書集合に分類する文書分類ステップと、該文書分類ステップにて分類された各部分文書集合からそれらの代表語セットを抽出する代表語抽出ステップと、任意の単語についてその関連語が記述された関連語辞書を用いて前記代表語抽出ステップにて抽出した各部分文書集合の代表語セットそれぞれについて関連語セットを抽出する関連語抽出ステップと、該関連語抽出ステップにて抽出した関連語セットと前記代表語抽出ステップで抽出した代表語セットと各部分文書集合に所属する文書に関する情報とをもとに個々の部分文書集合及び部分文書集合間の関連情報を生成する部分文書集合情報生成ステップと、前記文書分類ステップでの分類結果を前記部分文書集合情報生成部にて生成された情報と合わせて保存する分類結果保存ステップとを含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００５７】請求項２８の発明は、請求項２７に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記関連語抽出ステップにて抽出される関連語セットが、同義語、類義語、反対語のうちの少なくとも一つ以上の組合わせである文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００５８】請求項２９の発明は、請求項２７に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記関連語抽出ステップにて抽出される関連語セットが少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００５９】請求項３０の発明は、請求項２７に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記関連語抽出ステップにて抽出される関連語が少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットと代表語セットから反対語セットに対応する代表語を除いた単語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００６０】請求項３１の発明は、文書の内容に従って文書の分類を行う文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、文書データを入力する文書入力ステップと、前記文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析情報から文書データを多次元ベクトル空間で表現するための文書ベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データの分類を行う文書分類ステップとを含み、前記文書解析ステップにて抽出される特定の品詞を有する単語を、該特定の品詞の品詞情報に基づき、該特定の品詞の前後に抽出される一つ以上の単語と結合することにより生成される単語と置き換え、かつ該特定の品詞の品詞情報も適切に置き換える文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００６１】請求項３２の発明は、請求項３１に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書分類ステップにおいて統計手法としてクラスタリング法を用いることで文書データの分類を行う文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００６２】請求項３３の発明は、請求項３１または３２に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書解析ステップにおいて品詞が接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、単語及び品詞の置き換えを行う文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００６３】請求項３４の発明は、請求項３１ないし３３のいずれか１に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書解析ステップにおいて特定の品詞の単語が出現するまで単語の結合を続ける文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００６４】請求項３５の発明は、請求項３１ないし３４のいずれか１に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書解析ステップにおいて品詞が数詞接尾詞もしくは助数詞の単語について、該数詞接尾詞もしくは助数詞の単語に結合される複数の単語を削除し、前記文書分類ステップでは削除した単語の情報を用いない文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００６５】請求項３６の発明は、文書の内容に従って文書データ集合を分類する文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、文書データ集合を入力する文書入力ステップと、すべての文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析結果を記憶する文書解析結果記憶ステップと、前記文書解析ステップにて抽出された文書データの解析情報から前記文書データを多次元ベクトル空間で表現するためのベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成された文書ベクトル空間の各文書データのベクトルデータを記憶する文書ベクトルデータ記憶ステップと、指定される条件から文書データ集合の分類数を決定する分類数決定ステップと、前記文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データを前記指定された分類数の部分文書集合に分類する文書分類ステップと、該文書分類ステップで生成された分類結果を記憶する分類結果記憶ステップと、前記分類数決定ステップから前記分類結果記憶ステップまでの処理を繰り返し行うか否かの判定をおこなう繰り返し判定ステップと、前記文書ベクトルデータ記憶ステップと前記分類結果記憶ステップにて記憶された情報を用いて生成されたすべての部分文書集合間の関係情報を算出する部分文書集合間関係算出ステップと、該部分文書集合間関係算出ステップにて生成された部分文書集合間の関係情報を記憶する部分文書集合間関係記憶ステップとを含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００６６】請求項３７の発明は、請求項３６に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書分類ステップにて用いられる統計手法が非階層クラスタリング手法である文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００６７】請求項３８の発明は、請求項３６または３７に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記部分文書集合間関係算出ステップにて算出される関係が、類似関係と包含関係である文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００６８】請求項３９の発明は、請求項３８に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記部分文書集合間の関係は各部分文書集合から抽出される単語情報のみを用いて算出される文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００６９】

【発明の実施の形態】本発明の実施例の説明においては、自然言語で記述された１つ以上の文の集まりで、それが分類対象となる場合は、これを文書と言う。また、ひとつの文書の終端には、それが判別可能な文書終端記号が付置されているものとする。具体的な例をあげれば、公開特許公報や特定の新聞記事も文書であるし、それらから請求項や特定の１文を取り出したものであってもこれを文書と見なす。

【００７０】図１は本発明の請求項１，２，１４，１５，２７及び２８の発明に対応する実施例を説明するための文書分類装置のブロック構成図である。文書入力部１０１は、キーボード、ＯＣＲ装置、ハードディスク等の補助記憶装置等の入力手段が文書分類装置１００に直接に、または、ネットワーク経由で接続され、このような入力手段から文書や文書群を獲得し、文書データを入力するインターフェースである。図２は、文書データを入力する処理の一例を示すフローチャートである。

【００７１】図１における文書解析部１０２では、入力された文書それぞれに対し、自然言語解析を行い、単語やその品詞などを抽出する。さらに、文書内での単語の出現順序や、文書の作成者や作成日などの文書のメタ情報なども含めることができる。その後、文書群で出現した単語に対しユニークな単語ＩＤを付与し、文書内での単語出現回数を計数する。一例として、文書に対し形態素解析を適用することで、文書内の単語表記と品詞を抽出し、その結果をもとに文書群で出現したユニークな単語の表記、品詞、識別番号を抽出し、また各文書を抽出されたユニークな単語識別番号とその頻度で表現する例を示すこととし、そのフローチャートを図３に示す。

【００７２】例えば、図４（Ａ）に示す文書１と文書２に対し、形態素解析を適用すると図４（Ｂ）のような結果が得られる。図４（Ｂ）において各切り出された単語の下の数値はそれらの品詞を示しており、その対応表は図４（Ｃ）に示す。文書群が図４（Ａ）に示す２つの文書のみで構成されているとすると、文書群で出現したユニークな単語の表記、品詞、識別番号と各文書を単語識別番号とその頻度で表現した結果は図５（Ａ）～図５（Ｃ）のようになる。ただし、簡単のため品詞としては

名詞と未登録語のみを採用する。

【００７３】文書分類部１０３では、文書解析部１０２で生成された情報をもとに文書群の分類をおこなう。本発明では、分類手法は特に限定しないが、ここでは一例として、上記文書解析部１０２における実施例を継承して、各文書を文書群でユニークな単語の出現頻度のベクトルで表現し、これらのベクトルをもとにクラスタリング手法の１つであるｋ　ｍｅａｎｓ法を用いて文書分類を行う例を示すこととし、そのフローチャートを図６に示す。ここで、ベクトル間の類似度は０　１の間の実数、かつ最大類似度は１であるとする。

【００７４】図７（Ａ）に示す１５個の文書を図３及び図５に示すアルゴリズムを基に３つの部分文書集合に分類した結果を図７（Ｂ）に示す。ここで、品詞としては名詞と未登録語のみを採用し、またｋ　ｍｅａｎｓ法における類似測度は余弦測度であり、反復停止条件は繰返し回数５回としている。代表語抽出部１０４では、文書解析部１０２で生成した各文書の単語情報及び文書分類部で生成した部分文書グループに関する情報をもとに各部分文書集合における代表語セットを抽出する。

【００７５】本発明では、代表語の抽出方法を特に限定しないが、ここでは一例として、上記文書分類部における実施例を継承し、各部分文書集合においてそれらに所属する文書をひとつの仮想的な文書とみなした時の、文書群でユニークな単語の出現頻度が指定されたしきい値以上の単語をそれらの部分文書集合の代表語セットとする例を示すこととし、そのフローチャートを図８に示す。

【００７６】上記文書分類部における実施例の各部分文書集合について上記のフローチャートに従って求めた代表語セットを図９に示す。ここで、出現頻度のしきい値は２としている。

【００７７】関連語抽出部１０５では、代表語抽出部１０４にて抽出した各部分文書集合の代表語それぞれについて、関連語辞書を用いて関連語を抽出し、それらを各部分文書集合の関連語セットとする。関連語辞書としては、同義語辞書、広義語辞書、狭義語辞書、類義語辞書、反対語辞書、兄弟語辞書、上位概念語辞書、下位概念語辞書等を用いることができるが、ここでは一例として、上記代表語抽出部における実施例を継承し、任意の一つの関連語辞書を用いて各部分文書集合の関連語セットを求める例を示すこととし、そのフローチャートを図１０に示す。なお、複数の辞書を用いる場合には、前記処理を各辞書について繰り返し行えばよい。簡単のため関連語として同義語のみを扱うとして、前記代表語抽出部の実施例で求めた各代表語の同義語が図１１（Ａ）に示されるような場合、各部分文書集合の関連語セットは図１１（Ｂ）のように示される。

【００７８】部分文書集合情報生成部１０６では、文書解析部１０２で生成した各文書の単語情報、文書分類部

１０３で生成した部分文書グループに関する情報、代表語抽出部１０４で抽出した各部分文書集合の代表語セット、及び関連語抽出部１０５で生成した各部分文書集合の関連語セットを基に個々の部分文書集合及び部分文書集合間の関連情報を生成する。

【００７９】各部分文書集合固有の情報としては、代表語セットの集合、関連語セットの集合、及び各部分文書集合が多重分類を許す分類手法により生成されている場合は、代表語及び／または関連語を指定されるしきい値個数以上含む部分文書集合に所属する文書の部分集合等の情報を用いることができる。また、部分文書集合間の関連情報としては、部分文書集合間の代表語セット集合の積集合や和集合や差集合、関連語セットの集合の積集合や和集合や差集合、及び各部分文書集合が多重分類を許す分類手法により生成されている場合は、部分文書集合に所属する文書の積集合や和集合や差集合、代表語及び／または関連語を多く含む部分文書集合に所属する文書の部分集合間の積集合や和集合や差集合等の情報を用いることができる。

【００８０】ここでは一例として、上記関連語抽出部における実施例を継承し、文書部分集合が多重分類を許す分類手法により生成されているとしたときに、部分文書集合情報として、代表語セットの集合、関連語セットの集合、部分文書集合間の代表語セット集合の積集合と和集合と差集合、部分文書集合間の関連語セット集合の積集合と和集合と差集合を生成する例を示すこととし、そのフローチャートを図１２に示す。これらの情報により、特に部分文書集合間の類似性、関連性、及び包含関係などを把握することが可能になる。

【００８１】分類結果保存部１０７では、文書解析部１０２で生成した各文書の単語情報、文書分類部１０３で生成した部分文書グループに関する情報、代表語抽出部１０４で抽出した各部分文書集合の代表語セット、関連語抽出部１０５で生成した各部分文書集合の関連語セット、及び部分文書集合情報生成部１０６で生成した個々の部分文書集合及び部分文書集合間の関連情報を適切な形式で保存する。保存された関連情報は、出力部１０８からユーザの要求に応じて、または予め定められた条件に従って所定の出力手段に適宜出力される。

【００８２】図１３は本発明の請求項３，４，１６，１７，２９及び３０に対応する実施例を説明するための文書分類装置２００のブロック構成図である。なお、図１と同様の機能を有する部分には図１と同一の番号を付している。反意部分文書集合生成部２０１では、関連語抽出部１０５にて生成される関連語としてすくなくとも反対語が抽出されるとき、任意の文書部分集合が有する反対語が、自分を含む他のどの部分文書集合の代表語とも一致しない場合、この反対語を含む文書を文書群から抽出し、それを新しい部分文書集合とする処理をすべての部分文書集合について再帰的におこなう。

【００８３】ここでは一例として、上記実施例を継承して、関連語抽出部１０５にて反対語のみが抽出されることとし、各部分文書集合が有する反意語セットについてそれが自分を含む他の部分文書集合の代表語と一致するか否かを判定し、反意語がどの代表語とも一致しない場合、検索手法を用いて文書群からその反意語を含む文書を抽出し、それらを新しい部分文書集合とする例を示すこととし、そのフローチャートを図１４に示す。

【００８４】例えば、図７（Ａ）に示す文書群を分類した結果得られている図７（Ｂ）の部分文書集合３の代表語セットに着目してみる。この場合、代表語“商用”の反対語として、“無料、フリー”という単語が得られたとする。この場合、これらの単語はどの代表語とも一致せず、単語“無料”で文書群を検索した結果は該当０件であるが、単語“フリー”で検索した場合は、文書４、文書５、文書１２が検索される。これをあらたな部分文書集合とした場合、代表語として、“リナックス、フリー、ディストリビューション”を得ることができる。

【００８５】これにより文書群から任意の部分文書集合とは反対の意味を有する部分文書集合が文書分類部では生成されなかった場合にも、反対の意味を有する部分文書集合を生成することができるため、文書群からより広範囲な話題を抽出することが可能となる。

【００８６】請求項４，１７，３０の発明では、反対語からあらたな部分文書集合を求める際に、反対語を生成した代表語以外の部分文書集合の代表語も合わせて部分文書集合を求めることにより、より対象の部分文書集合とは反対の意味をもつ部分文書集合を生成することが可能となるが、基本的な処理は上記実施例と同様の処理で求めることができる。すなわち、例えば、図１４に示すフローチャートにおいて反対語を用いて文書群を検索するステップを反対語と反対語を生成した代表語以外の部分文書集合の代表語を組合わせた論理式を用いればよい。

【００８７】図１５は、本発明の請求項５～９，１８～２２及び３１～３５に対応する実施例を説明するための文書分類装置のブロック構成図である。文書入力部３０１は、キーボード、ＯＣＲ装置、ハードディスク等の補助記憶装置等の入力手段が文書分類装置３００に直接に、または、ネットワーク経由で接続され、このような入力手段から文書や文書群を獲得し、文書データを入力するインターフェースである。この際、各文書データを一意に識別するために、例えばユニークな数などの、識別子を各文書に割り当てる。

【００８８】文書解析部３０２では、入力された文書それぞれに対し形態素解析を適用し、各文書を構成する単語を品詞情報等とともに抽出する。この際、抽出した単語を識別するために、抽出した単語のうちユニークな表記を持つものについては、ユニークな識別子を付置しておく。さらに、形態素解析の結果得られる単語のうち指定される品詞をもつ単語について、その前後の単語と適切に組合わせた単語と置き換え、かつ品詞もまた適切なものに置き換える処理を施す。例として、品詞が接頭詞全般、接尾詞全般、及び助数詞である単語について前記の結合及び置き換え処理を行う動作を説明する。

【００８９】まず、本例では、前記の結合および置き換え処理を品詞が、１．接頭詞全般、２．数詞接尾詞以外の接尾詞全般、３．数詞接尾詞もしくは数助詞の場合別に以下のような規則でおこなうこととする。ただし、本発明における結合及び置き換え処理の規則はこれらに限定するものではない。

【００９０】

```
○接頭詞全般
もし｛対象単語の品詞が接頭詞である｝ならば｛
  計数用変数：ｉに１を代入する
  繰り返す｛
    対象単語の先頭に対象単語よりｉ回前に抽出された単語を結合させる
    もし｛i回前に抽出されている単語の品詞が分類時使用品詞である}ならば{
      繰り返しループを抜ける
    }
    さもなくば｛
      ｉを１増加する
    }
  }
  対象単語の品詞を変更する
}
```

【００９１】

```
○数詞接尾詞以外の接尾詞全般
もし｛対象単語の品詞が数詞接尾詞以外の接尾詞である｝ならば｛
  計数用変数：ｉに１を代入する
  繰り返す｛
```

```
        対象単語の終端に対象単語よりｉ回後に抽出された単語を結合させる
        もし｛i回前に抽出されている単語の品詞が分類時使用品詞である｝ならば{
          繰り返しループを抜ける
        }
        さもなくば｛
          ｉを１増加する
        }
      }
      対象単語の品詞を変更する
    }
```

【００９２】

```
    ○数詞接尾詞もしくは助数詞
    もし｛対象単語の品詞が数詞接尾詞もしくは助数詞である｝ならば｛
      繰り返す｛
        もし｛対象単語の直前に抽出されている単語の品詞が数詞である｝ならば｛
          対象単語の先頭に対象単語の直前に抽出された単語を結合させる
          対象単語のｉ回前に抽出された単語を削除する
        }
        さもなくば｛
          繰り返しループを抜ける
        }
      }
      対象単語の品詞を変更する
    }
```

【００９３】図１６に示す６つの文書データを分類対象文書データとし、この文書データに対して形態素解析を適用し、単語及びそれらの品詞を抽出したものを図１７に示す。ただし、本発明では形態素解析系については特に規定しない。また、分類時使用品詞を普通名詞、サ変名詞、固有名詞、数詞、形容詞、接頭詞全般、接尾詞全般、助数詞賭した場合の文書データの解析結果を図１８に示す。

【００９４】図１８に示されている結果において、品詞が接頭詞全般、接尾詞、もしくは数助詞である単語に対し前記規則に従い、結合・置き換え処理を施した結果を図１９に示す。例えば、文書１における｛千葉［普通名詞］、氏［固有名詞接尾詞］｝という文字列は、数詞接尾詞以外の接尾詞全般の規則を用いて、｛千葉［普通名詞］、千葉氏［固有名詞］｝という文字列になり、また｛１［数詞］、９［数詞］、５［数詞］、０［数詞］、年［助数詞］｝という文字列は、数詞接尾詞もしくは助数詞の規則を用いて、｛１９５０年［普通名詞］｝という文字列になる。

【００９５】文書ベクトル空間生成部３０３では、前記文書解析部にて抽出された各文書データの単語情報をもとに文書データをベクトル表現するための空間を生成する。例として、前記文書解析部での例をもとに、文書データ全体でユニークな単語の頻度により文書ベクトル空間を生成することとする場合の各文書データのベクトル表現を生成する動作を説明する。ただし、本発明では、ベクトル空間生成手法はこれに限定するものではなく、例えば、全単語の線形変換によりベクトル空間を生成することもできる。

【００９６】図１８及び図１９に示す文書解析結果からユニークな単語を抽出し、各文書での該当単語の頻度を計数し、それらの結果を、単語を列方向に、文書データを行方向に付置することで、行列表現したものをそれぞれ図２０と図２１に示す。これら行列において、列ベクトルが各文書データのベクトルデータとなる。

【００９７】文書分類部３０４では、前記文書ベクトル空間生成部にて生成された文書データベクトルを統計手法を用いることで幾つかの集合に分類する。出力部３０５では、文書分類部３０４で分類された文書データベクトルの集合をユーザの要求に応じてまたは予め定められた条件に従って所定の出力手段に適宜出力する。文書分類部３０４における統計処理は様々なものが利用可能であるが、請求項５の発明ではアルゴリズムの簡潔さやパラメータの有無等の理由からクラスタリング手法を用いることに限定している。例として、前記文書ベクトル空間生成部での例をもとに、クラスタリング手法を用いて文書ベクトルを分類する動作を説明する。

【００９８】ここでは、クラスタリング手法の１つであるＷａｒｄ法を用いることとし、また類似測度は標準化ユークリッド距離測度を使用する。なお、クラスタリング手法に関しては、“多変量解析入門（森北出版）”に詳しい。図２０及び図２１に示されている文書データ

に対し、Ｗａｒｒｄ法を適用した結果を図２２と図２３に示す。ここで、図２０は前記結合・置き換えの処理を適用した結果で文書ベクトル空間を構成したデータであり、図２１は結合・置き換え処理を適用していない結果で文書ベクトル空間を構成したデータである。また、図２２と図２３の図中の数値は各クラスタ間の距離である。

【００９９】図２２及び図２３の結果を比較した場合、文書４の位置の差異が非常に特徴的であり、結合・置き換えの処理を適用した場合は、文書４は文書２や文書５と類似していると判断され、結合・置き換えの処理を適用しない場合は、文書４は文書１や文書６と類似していると判断される。主観的な語彙の適合度などから判断して文書４は｛文書２、文書５｝の集合よりも｛文書１、文書３、文書６｝の集合に含まれる方が適切であると思われる。従って、この結果から、結合・置き換えの処理を適用することにより、より質の高い文書ベクトル空間を構成でき、この文書ベクトル空間で分類処理をおこなうことで、質の高い文書分類結果を得ることができる。

【０１００】図２４は本発明の請求項１０～１３，２３～２６及び３６～３９に対応する実施例を説明するための文書分類装置のブロック構成図である。文書入力部４０１は、キーボード、ＯＣＲ装置、ハードディスク等の補助記憶装置による入力手段が文書分類装置４００に直接に、または、ネットワーク経由で接続され、このような入力手段から文書や文書群を獲得し、文書データを入力するインターフェースである。この際、各文書データを一意に識別するために、例えばユニークな数などの、識別子を各文書に割り当てる。

【０１０１】文書解析部４０２では、入力された文書それぞれに対し形態素語解析を適用し、各文書を構成する単語を品詞情報等とともに抽出する。この際、抽出した単語を識別するために、抽出した単語のうちユニークな表記を持つものについては、前記文書データと同様にユニークな識別子を付置しておく。例として、文書データに対し形態素解析を適用し、文書データ全体で表記と品詞がユニークである単語を同定し、それらに一意な識別番号を付与するとともに、各文書データを、それを構成する単語の識別番号とその出現頻度を表現するための擬似コードを図２５に示す。なお、本発明では、形態素解析系は必要な情報を抽出できるものであれば、どのようなものでもよい。

【０１０２】文書解析結果記憶部４０３では、文書解析部４０２にて抽出された文書データの形態素解析結果を適切な形式で記憶する。文書ベクトル空間生成部４０４では、文書解析部４０２にて抽出された各文書データの単語情報をもとに文書データをベクトル表現するための空間を生成する。例として、文書解析部４０２での例をもとに、文書データ全体でユニークな単語の正規化された頻度により文書ベクトル空間を生成する場合の、各文書データのベクトル表現を生成する擬似コードを図２６に示す。ただし、本発明では、ベクトル空間生成手法はこれに限定するものではなく、例えば、特異値分解などを使用して全単語の線形変換によりベクトル空間を生成することもできる。

【０１０３】文書ベクトルデータ記憶部４０５では、文書ベクトル空間生成部４０４にて生成された文書データベクトルを適切な形式で記憶する。分類数決定部４０６では、繰り返し文書分類を行う際の分類数を決定する（分類数を定数×繰返し数とした場合の擬似コードを図２７に含む）。文書分類部４０７では、文書ベクトル空間生成部４０４にて生成された文書データベクトルを統計手法を用いることで分類数決定部集合に分類する。

【０１０４】統計処理は様々なものが利用可能であるが、請求項１１の発明ではアルゴリズムの簡潔さやクラスタ数の変化により分類構造が動的に変化する特性等から非階層クラスタリング手法を用いることに限定している。例として、クラスタ数を繰返し数と定数Ｎを乗じた数としてクラスタリング手法を用いて文書ベクトルを分類する擬似コードを図２７に示す。ここでは、クラスタリング手法の１つであるｋ　ｍｅａｎｓ法を一部変更したもの用いることとし、また類似測度は余弦測度を使用する。なお、クラスタリング手法に関しては、“多変量解析入門（森北出版）”に詳しい。

【０１０５】文書分類結果記憶部４０８では、文書分類部４０７で生成される文書分類結果を適切な形式で記憶する。繰り返し判定部４０９では、繰り返し文書分類をおこなう際の繰り返しを継続するか否かの判定を行う（繰り返し判定を指定された最大数を限度とした場合の擬似コードを図２７に含む）。部分文書集合間関係算出部４１０では、文書分類結果記憶部４０８に記憶されている複数の部分文書集合間の関係情報を、文書解析結果記憶部４０３と文書ベクトルデータ記憶部４０５にて記憶されている種々の文書データに関する情報を用いて算出する。例として、部分文書集合間の類似関係と包含関係を文書データ及び／または文書データを構成する単語情報で算出する動作を説明する。

【０１０６】まず、部分文書集合間の類似関係と包含関係を文書データで表現するための定式化を行う。文書分類結果記憶部４０８に記憶されている複数の部分文書集合はユニークな識別番号が付与されているものとする。第ｍ番目の部分文書集合の特性ベクトル：Ｖｍを以下のように定義する。

【０１０７】・Ｖｍの次元数は全文書データ数に等しい
・Ｖｍの各要素はそれぞれ１つの文書データに対応し、重複はない。
・要素ｉに対応する文書データと部分文書集合との類似度が閾値以上の場合、要素ｉは１となる。
・要素ｉに対応する文書データと部分文書集合との類似度が閾値未満の場合、要素ｉは０となる。

【０１０８】上記定義を用いて、第m番目の部分文書集合と第ｎ番目の部分文書集合の関係：ＲｍｎとＲｎｍを以下のように定義する。

（１）Ｒｍｎ＝＜Ｖｍ，Ｖｎ＞／＜Ｖｍ，Ｖｍ＞

（２）Ｒｎｍ＝＜Ｖｍ，Ｖｎ＞／＜Ｖｎ，Ｖｎ＞

ただし、＜，＞は内積を示す。

【０１０９】上記のＲｍｎとＲｎｍの値により、部分文書集合間の類似関係と包含関係を算出することが可能となる。図２８はＲｍｎとＲｎｍの値による幾何学的解釈を示したものである。すなわち、Ｒｍｎが１に近い場合は、部分文書集合ｍは部分文書集合ｎに包含されているといえる。また、ＲｍｎとＲｍｎが両方１に近いほど部分文書集合ｍと部分文書集合ｎは類似しているものといえる。さらに、（Ｒｍｎ，Ｒｎｍ）がＲｍｎ＝Ｒｎｍの直線に近いほど、同じ程度の割合で相互に文書データを包含していることなども読み取れる。

【０１１０】次に、部分文書集合間の類似関係と包含関係を文書データを構成する単語の出現頻度情報で表現するための定式化をおこなう。第m番目の部分文書集合の特性ベクトル：Ｗｍを以下のように定義する。

【０１１１】Ｗｍの次元数は全文書データでユニークな単語数に等しい。Ｗｍの各要素はそれぞれユニークな単語に対応し、重複はないＷｍの第Ｉ番目の要素値をｗｍ（ｉ）と示す。部分文書集合との類似度が閾値以上の文書すべてにおける、要素ｉに対応する単語の出現頻度（出現回数）を要素ｉの要素値とする。

【０１１２】上記定義を用いて、第m番目の部分文書集合と第ｎ番目の部分文書集合の関係：Ｒ’ｍｎとＲ’ｎｍを以下のように定義する。

（３）R’mn = Σf(wm(k),wn(k)) /Σf(wm(k),wm(k))

（４）R’mn = Σf(wm(k),wn(k)) /Σf(wn(k),wn(k))

（５）f(wm(k),wn(k)) = 0 for wm(k)×wn(k) = 0= wm(k)×(a + b / | wn(k) - wn(k) | + 1) for forwm(k)×wn(k) != 0

ただし、a,bは定数で、a + b = 1，a,b >= 0

【０１１３】上記のＲ’ｍｎとＲ’ｎｍの値を用いても図２８に示すＲｍｎとＲｍｎの関係と同様の解釈ができ、したがって、部分文書集合間の類似関係と包含関係を算出することが可能となる。さらに、Ｒ’ｍｎとＲ’ｎｍを用いて部分文書集合の関係を定義する場合、文書データのレベルでは得ることのできない関係を得ることが可能になるとともに、例えば内容は一致してても、分析対象の文書データが異なっている場合にも部分文書集合間の関係を算出することが可能となる。また、部分文書集合間関係記憶部４１１では、部分文書集合間関係算出部４１０にて生成された部分文書集合間の関係情報を適切な形式で記憶する。また、出力部４１２は、部分文書集合間関係記憶部４１１で記憶された関係情報をユーザの要求に応じて、または予め定められた条件に従って出力手段に適宜出力する。

【０１１４】

【発明の効果】請求項１，１４及び２７の発明によれば、生成された部分文書集合それぞれの代表語セットを抽出し、さらにそれら代表語それぞれについて関連語を求め、これらの情報をもとに各部分文書集合および部分文書集合間の関連情報を生成することで、部分文書集合の分析に有効な情報を提供することができる。

【０１１５】請求項２，１５及び２８の発明によれば、関連語として同義語、類義語、反対語のすくなくとも一つ以上の組合わせを用いることで主に類似性に関す情報を提供することができる。

【０１１６】請求項３，４，１６，１７，２９及び３０の発明によれば、各部分文書集合の代表語セットの関連語として反対語を用い、反対語が自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、その反対語を含む文書を文書集合から抽出し、それを新たな部分文書集合とすることで、文書集合からより多くの分析情報を抽出することができる。

【０１１７】請求項５，１８及び３１の発明によれば、分類対象文書に形態素解析し、得られた解析結果をもとに分類対象文書を幾つかの文書集合に分類する文書分類装置において、形態素解析の結果得られる単語のうち指定される品詞をもつ単語について、その前後の単語と適切に組合わせた単語と置き換え、かつ品詞もまた適切なものに置き換える処理を施すことによって、高品位な文書ベクトル空間を構成し、この文書ベクトル空間で統計処理を用いて文書分類を行うことで高品質な文書分類結果を得ることができる。

【０１１８】請求項６，１９及び３２の発明によれば、文書分類をおこなうための統計手法として、クラスタリング手法を用いることで、簡便に高品質な文書分類結果を得ることができる。

【０１１９】請求項７，２０及び３３の発明によれば、分類対象文書に形態素解析を適用することで抽出される単語の中で、特に、品詞が、接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、適切な結合処理を施こすことで、高品質な文書ベクトル空間を得ることができる。

【０１２０】請求項８，２１及び３４の発明によれば、単語の結合処理において特定の品詞の単語が出現するまで単語の結合を続けることによって新たな単語を生成することで、高品質な文書ベクトル空間を得ることができる。

【０１２１】請求項９，２２及び３５の発明によれば、単語の結合処理において、品詞が数詞接尾詞もしくは助数詞の単語について、結合される複数の単語を削除し、文書ベクトル空間を生成する際にはそれらの単語の情報は用いないことで、高品質な文書ベクトル空間を得ることができる。

【０１２２】請求項１０，２３及び３６の発明によれ

ば、文書のベクトル空間モデルを用い、生成する部分文書集合の数をパラメータとして繰り返し文書分類処理をおこなうことで、多数の部分文書集合を生成し、さらに生成された多数の文書集合について相互の関係を算出することで、文書集合の構造の把握を支援しうる情報を生成する文書分類装置を提供することができる。

【０１２３】請求項１１，２４及び３７の発明によれば、上記目的に加え、文書分類をおこなう統計手法として、非階層クラスタリング手法を用いることで、簡便に多数の部分文書集合を生成することができる。

【０１２４】請求項１２，２５及び３８の発明によれば、上記目的に加え、生成された多数の文書集合について相互の関係として、類似関係と包含関係を算出することで、容易に文書集合の構造の把握しうる情報を提供することができる。

【０１２５】請求項１３，２６及び３９の発明によれば、上記目的に加え、生成された多数の文書集合が有する情報のうち、単語に関する情報のみを用いて相互の関係を算出することで、汎用性・再利用性の高い関係情報を算出することができる。

【図面の簡単な説明】

【図１】　本発明の請求項１，２，１４，１５，２７及び２８の発明に対応する実施例を説明するための文書分類装置のブロック構成図である。

【図２】　文書データを入力する処理の一例を示すフローチャートである。

【図３】　文書に対し形態素解析を適用する処理の一例を示すフローチャートである。

【図４】　形態素解析の適用例について説明するための図である。

【図５】　形態素解析の適用結果の一例について説明するための図である。

【図６】　文書解析部で生成された情報をもとに文書群の分類を行う処理の一例を示すフローチャートである。

【図７】　文書の部分文書集合への分類を説明するための図である。

【図８】　代表語の抽出の処理の一例を示すフローチャートである。

【図９】　図８に示すフローチャートに従って求めた代表語セットの一例を示す図である。

【図１０】　各部分文書集合の代表語セットのそれぞれについて関連語辞書を用いて関連語を抽出する処理の一例を示すフローチャートである。

【図１１】　代表語抽出部で求め各代表語の同義語及び各部分文書集合の関連語セットの一例を示す図である。

【図１２】　抽出または生成した代表語セット及び関連語セットを共に個々の部分文書集合及び部分文書集合間の関連情報を生成する処理の一例を示すフローチャートである。

【図１３】　本発明の請求項３，４，１６，１７，２９及び３０に対応する実施例を説明するための文書分類装置のブロック構成図である。

【図１４】　反対語を含む文書を抽出して新しい部分文書集合とする処理の一例を示すフローチャートである。

【図１５】　本発明の請求項５～９，１８～２２及び３１～３５に対応する実施例を説明するための文書分類装置のブロック構成図である。

【図１６】　分類対象文書データの例を示す図である。

【図１７】　図１６に示す文書データに形態素解析を適用して単語及び品詞を抽出した例を示す図である。

【図１８】　文書データの解析結果の一例を示す図である。

【図１９】　文書データの解析結果の他の例を示す図である。

【図２０】　文書データを行方向に位置することで行列表現した例を示す図である。

【図２１】　文書データを行方向に位置することで行列表現した他の例を示す図である。

【図２２】　図２０の文書データに対しＷａｒｒｄ法を適用した結果を示す図である。

【図２３】　図２１の文書データに対しＷａｒｒｄ法を適用した結果を示す図である。

【図２４】　本発明の請求項１０～１３，２３～２６及び３６～３９に対応する実施例を説明するための文書分類装置のブロック構成図である。

【図２５】　文書データの単語の識別番号とその出現頻度を表現するための擬似コードの一例を示す図である。

【図２６】　各文書データのベクトル表現を生成する擬似コードの一例を示す図である。

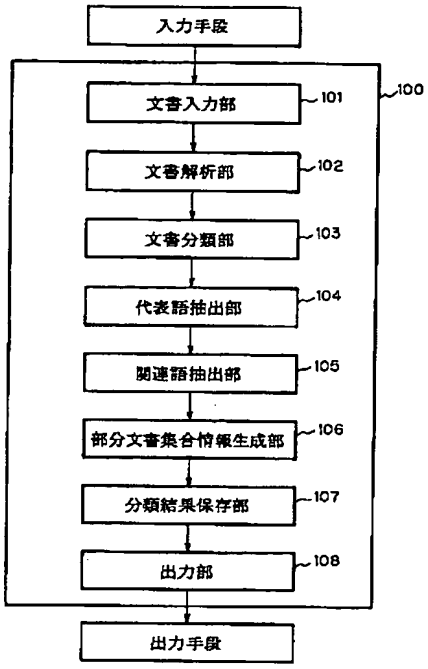
【図２７】　クラスタリング手法を用いて文書ベクトルを分類する擬似コードの一例を示す図である。

【図２８】　ＲｍｎとＲｎｍの値による幾何学的解釈を示したものである。

【符号の説明】

１００，２００，３００，４００…文書分類装置、１０１，３０１，４０１…文書入力部、１０２，３０２，４０２…文書解析部、１０３，３０４，４０７…文書分類部、１０４…代表語抽出部、１０５…関連語抽出部、１０６…部分文書集合情報生成部、１０７…分類結果保存部、１０８，３０５，４１２…出力部、２０１…反意部分文書集合生成部、３０３，４０４…文書ベクトル空間生成部、４０３…文書解析結果記憶部、４０５…文書ベクトルデータ記憶部、４０６…分類数決定部、４０８…文書分類結果記憶部、４０９…繰り返し判定部、４１０…部分文書集合間関係算出部、４１１…部分文書集合間関係記憶部。

【図１】



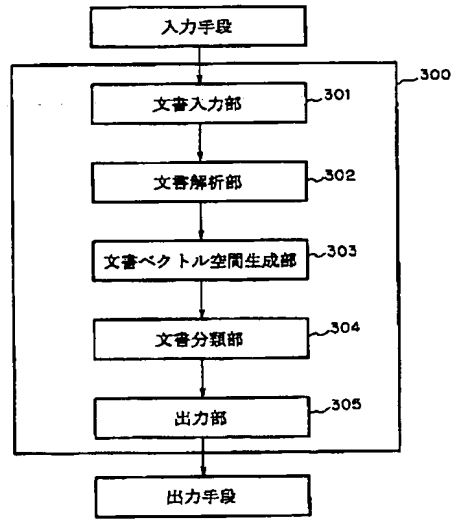
【図２】

```
文書群の文書数用変数：Ｎを定義し、文書数を代入する。
大きさＮの文書配列変数：ｄｏｃ［ｎ］（ｎ＝１？ Ｎ）を定義する。
繰り返し変数：ｉに１を代入する。
ｗｈｉｌｅ（ｉがＮ以下である）{
  ｄｏｃ［ｉ］にｉ番めの文書を代入する。
  ｉを１増加する。
}
```

【図９】

部分文書集合１	部分文書集合２	部分文書集合３	部分文書集合４
リナックス カーネル ＯＳ 可能 オペレーティングシステム 配布	リナックス 日本語 環境 ディストリビューション Ｖｉｎｅ	リナックス 商用 アプリケーション 開発 利用	インストール

【図１５】



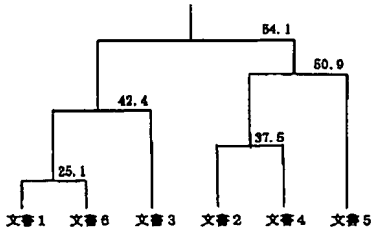
【図１６】

文書識別番号	文書内容
1	千葉氏は1950年にフランス共和国に移住した。
2	近年フランスに移住を希望する人が増加している。
3	千葉商会にチーズを1950個発注した。
4	フランス産の場合、希望小売価格は1950円である。
5	千葉産のチーズの価格はフランス産に比べると格段に安い。
6	そして、1950年当時、まだ星を沢山見ることが出来た。

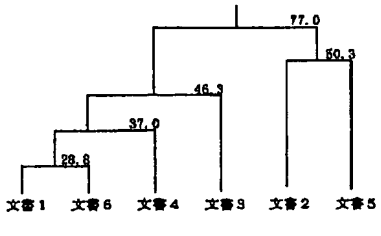
【図３】

```
要素が大きさが可変長の単語識別番号保持用配列変数：ｉＷｏｒｄＩＤ〔〕、同じく可変長の大きさの単語頻度保
持用配列変数：ｉＷｏｒｄＦｒｅｑ〔〕、単語識別番号数保持用変数：ｉＷｏｒｄＳｉｚｅ、および文書識別番号
保持用変数：ｉＤｏｃＩＤをもつ文書構成要素型：ＴＹＰＥ＿ＤｏｃＰａｒｔｓを定義する。

大きさが可変長の単語表記用配列変数：ｗｏｒｄＮｏｔｅ〔〕を定義する。
大きさが可変長の単語品詞用配列変数：ｗｏｒｄＣｌａｓｓ〔〕を定義する。
大きさが可変長の単語識別番号用配列変数：ｗｏｒｄＩＤ〔〕を定義する。
文書群のユニークな単語の単語数用変数：ｕＷｏｒｄｓを定義する。
大きさがＮのＴＹＰＥ＿ＤｏｃＰａｒｔｓ型配列変数：ｄｏｃＰａｒｔｓ〔ｎ〕（ｎ＝1? Ｎ）を定義する。
繰り返し変数：ｉにｌを代入する。
繰り返し変数：ｊにｌを代入する。
ｕＷｏｒｄｓに０を代入する。
ｗｈｉｌｅ（ｉがＮ以下である）｛
  ｄｏｃＩＤ〔ｉ〕に形態素解析を適用する。
  抽出された単語数を変数：Ｍに代入する。
  繰り返し変数：ｊにｌを代入する。
  ｄｏｃＰａｒｔｓ〔ｉ〕のｉＤｏｃＩＤにｉを代入する。
  ｄｏｃＰａｒｔｓ〔ｉ〕のｉＷｏｒｄＳｉｚｅにＭを代入する。
  ｄｏｃＰａｒｔｓ〔ｉ〕のｉＷｏｒｄＩＤとｉＷｏｒｄＦｒｅｑの大きさをＭとし、０を代入する。
  ｗｈｉｌｅ（ｊがＭ以下である）｛
    ｊ番目に抽出された単語の表記を変数：ｗｎに代入する。
    ｊ番目に抽出された単語の品詞を変数：ｗｃに代入する。
    繰り返し変数：ｋにｌを代入する
    ｗｈｉｌｅ（ｋがｕＷｏｒｄｓ以下である）｛
      ｉｆ（ｗｎがｗｏｒｄＮｏｔｅ〔ｋ〕と一致しかつｗｃがｗｏｒｄＣｌａｓｓ〔ｋ〕と一致する）｛
        繰り返しループから抜ける。
      ｝
      ｅｌｓｅ｛
        ｋを１増加する。
      ｝
    ｝

    ／／　ｗｎがすでに出現している単語の場合、ｋはｕＷｏｒｄｓ以下である
    ｉｆ（ｋがｕＷｏｒｄｓより大きい）｛
      ｕＷｏｒｄｓを１増加する。
      ｗｏｒｄＮｏｔｅ、ｗｏｒｄＣｌａｓｓ、ｗｏｒｄＩＤを大きさを１大きくする。
      単語ｗｎはユニークな単語であるので、ｗｏｒｄＮｏｔｅ〔ｕＷｏｒｄｓ〕にｗｎを代入する。
      ｗｏｒｄＣｌａｓｓにｗｃ〔ｕＷｏｒｄｓ〕にｗｃを代入する。
      ｗｏｒｄＩＤ〔ｕＷｏｒｄｓ〕にｕＷｏｒｄｓを代入する。
    ｝
    ｄｏｃＰａｒｔｓ〔ｉ〕のｉＷｏｒｄＩＤ〔ｊ〕にｋを代入する。
  ｝
  ｄｏｃＰａｒｔｓ〔ｉ〕のｉＷｏｒｄＩＤを昇順にソートし、並び替えの順序情報を順序配列変数：ｍａｐに代
入する。
  順序配列変数：ｍａｐをもとにｄｏｃＰａｒｔｓ〔ｉ〕のｉＷｏｒｄＦｒｅｑを並び替える。

｝
```

【図２２】



【図２３】



【図４】

(A)

文書1：
リナックス というのは自由に再配布することのできる，
独立した Unix 系オペレーティングシステム(OS)のことです．

文書2：
リナックス は完全にスクラッチから書き起こされた OS であり，
そのおかげであらゆる意味における「再配布」が可能なのです．

(B)

文書1：
リナックス と いう の は 自由 に 再 配布 する こと の できる 、
1 5 2 5 5 1 5 14 1 7 1 5 2 10
独立 し た Unix 系 オペレーティングシステム （ OS ） の こと です 。
1 7 7 8 14 1 10 8 10 5 1 7 10

文書2：
リナックス は 完全 に スクラッチ から 書き 起こさ れ た OS で あり 。
1 5 1 5 8 5 2 2 7 7 8 7 2 10
その おかげ で あらゆる 意味 に おける 「 再 配布 」 が 可能 な のです 。
11 1 5 11 1 5 8 10 1 1 10 12 1 5 8 10

(C)

1：名詞、2：動詞
3：形容詞、4：形容動詞
5：助詞、6：副詞
7：助動詞、8：未登録語
9：数詞、10：記号
11：連体詞、12：接続詞
13：感動詞、14：接辞

【図５】

(A)

識別番号（ｗｏｒｄＩＤ）	単語表記（ｗｏｒｄＮｏｔｅ）	品詞（ｗｏｒｄＣｌａｓｓ）
1	リナックス	1
2	自由	1
3	再	1
4	配布	1
5	独立	1
6	UNIX	8
7	オペレーティングシステム	1
8	ＯＳ	8
9	完全	1
１０	スクラッチ	8
１１	おかげ	1
１２	意味	1
１３	おける	8
１４	可能	1

(B)

文書１

単語識別番号（ＩＷｏｒｄＩＤ）	単語頻度（ｉＷｏｒｄＦｒｅｑ）
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1

(C)

文書２

単語識別番号（ＩＷｏｒｄＩＤ）	単語頻度（ｉＷｏｒｄＦｒｅｑ）
1	1
9	1
１０	1
8	1
１１	1
１２	1
１３	1
3	1
4	1
１４	1

【図６】

```
//　文書内のユニークな単語の頻度のベクトルの生成
文書内のユニークな単語の頻度のベクトル用の大きさがＮ×ｕＷｏｒｄｓの２次元配列変数：ＤｏｃＶｅｃｔｏｒ
［Ｎ］［ｕＷｏｒｄｓ］を定義する。
繰り返し変数：ｉに１を代入する。
ｗｈｉｌｅ（ｉがＮ以下である）｛
  繰り返し変数：ｊに１を代入する。
  繰り返し変数：ｋに１を代入する。
  ｗｈｉｌｅ（ｊがｕＷｏｒｄｓ以下である）｛
    ｉｆ（ｊがｄｏｃＰａｒｔｓ［ｉ］のｉＷｏｒｄＩＤ［ｋ］と等しい）｛
      ＤｏｃＶｅｃｔｏｒ［ｉ］［ｊ］にｄｏｃＰａｒｔｓ［ｉ］のｉＷｏｒｄＦｒｅｑ［ｋ］を代入する。
      変数ｋを１増加する。
    ｝
    ｅｌｓｅ｛
      ＤｏｃＶｅｃｔｏｒ［ｉ］［ｊ］を代入する。
    ｝
    変数ｊを１増加する。
  ｝
  変数ｉを１増加する。
｝
//　書記クラスタ重心の設定
分類数用変数：Ｃを定義し、分類数を代入する。（ＣはＮ以下でなければならない）
クラスタ重心ベクトル用の大きさがＣ×ｕＷｏｒｄｓの２次元配列変数：ＣＶｅｃｔｏｒ［Ｃ］［ｕＷｏｒｄｓ］
を定義する。
繰り返し変数：ｉに１を代入する。
ｗｈｉｌｏ（ｉがＣ以下である）｛
  繰り返し変数：ｊに１を代入する。
  ｗｈｉｌｅ（ｊがｕＷｏｒｄｓ）以下である）｛
    ＣＶｅｃｔｏｒ［ｉ］［ｊ］にＤｏｃＶｅｃｔｏｒ［ｉ］［ｊ］を代入する。
  ｝
｝
//　精練化処理
//　多重分類の場合、ＤｏｃＰｏｓの第２次元が１以上となる。
各文書の所属クラスタ番号用の大きさがＮ×可変である２次元配列変数：ＤｏｃＰｏｓ［Ｎ］［］を定義する。こ
の場合、ＤｏｃＰｏｓ［Ｎ］［１］である。
最大類似度用変数：ｍａｘＳｉｍを定義する。
類似度用変数：ｓｉｍを定義する。
ｗｈｉｌｅ（反復停止条件を満たしていない）｛
  繰り返し変数：ｉに１を代入する。
  ｗｈｉｌｅ（ｉがＮ以下である）｛
    繰り返し変数：ｊに１を代入する。
    ｍａｘＳｉｍに０を代入する。
    ｗｈｉｌｅ（ｊがＣ以下である）｛
      ｉ番目の文書；ＤｏｃＶｅｃｔｏｒ［ｉ］とｊ番目のクラスタ重心ＣＶｅｃｔｏｒ［ｊ］との類似度を計算
      し、ｓｉｍに代入する。
      ｉｆ（ｓｉｍがｍａｘＳｉｍ以上である）｛
        ＤｏｃＰｏｓ［ｉ：１］にｊを代入する。
        ｍａｘＳｉｍにｓｉｍを代入する。
      ｝
    ｝
    ＣＶｅｃｔｏｒ［ＤｏｃＰｏｓ［ｉ］］をＣＶｅｃｔｏｒ［ＤｏｃＰｏｓ［ｉ］］とＤｏｃＶｅｃｔｏｒ［ｉ］
    の平均とする。
  ｝
｝
```

【図７】

(A)

文書１：リナックス というのは自由に再配布することのできる，独立した Unix 系オペレーティングシステム(OS)のことです．
文書２：リナックス は完全にスクラッチから書き起こされた OS であり，そのおかげであらゆる意味における「再配布」が可能なのです．
文書３：リナックス は本来 OS の中核となるカーネル(kernel)だけを指す名称ですが，リナックス カーネルベースのシステム全体をさして 「リナックス」と表現することもあります．
文書４：いくつかのディストリビューションのベンダーは，インストール用ソフトウェアを含めたディストリビューションを，インターネット上でフリーで入手できるようにしています．
文書５：リナックス カーネルのソースコードは常にフリーに入手可能でなくてはなりません．
文書６：リナックス がインターネットで自由に手に入るようになり，多くのベンダーは〝ディストリビューション〟と呼ばれるパッケージ化されたバージョンのリナックス を作りました．
文書７：リアルタイムなオペレーティングシステムを必要とする人々は，リナックス・カーネルを元にした小さなリアルタイム・カーネルに移行することを可能 にしました．
文書８：望めばリナックス に対して金銭をやり取りすることも可能ですが，その場合もリナックス の再配布を制限することはできません．
文書９：ソフトウェア開発，ネットワーク利用(インターネット，イントラネット，SOHO などなど)，そしてエンドユーザのプラットフォームとしてリナックス は高価な商用UNIX システムに取って代わりつつあるのです．
文書１０：Vine リナックス は，日本語版のインストーラによって、特に初心者のインストール時の不安を減らしています。
文書１１：海外の商用アプリケーションは RedHat リナックス を前提に開発されていることが多く、その多くをそのまま利用することができます。
文書１２：商用フォントやアプリケーションを含めた 製品版が提供されても，Vine リナックス がフリーであることに変わりはありません 。
文書１３：開発は，Slackware や RedHat リナックス に対応した日本語環境 add-on として定評のある，PJE のメンバーを中心に進められています．
文書１４：Vine リナックス は，使いやすい日本語環境を提供する リナックス・ディストリビューションです．
文書１５：インストールの直後から、快適な日本語環境で作業をできるように、様々な配慮を行っております。

(B)

部分文書集合１	部分文書集合２	部分文書集合３	部分文書集合４
文書７	文書１４	文書１２	文書１５
文書５	文書１３	文書１１	文書４
文書３	文書１０	文書９	
文書１	文書８		
文書２	文書６		

【図８】

```
各部分文書集合における代表語セット用の大きさがＣ×可変の２次元配列変数：ｔｙｐｉｃａｌＷｏｒｄｓ［Ｃ］
［］を定義する。
各部分文書集合における代表語セット数用の大きさがＣの配列変数：ｔｙｐｉｃａｌＷｏｒｄｓＳｉｚｅ［Ｃ］を
定義する。
代表語判定をおこなうためのしきい値用変数：ｔｈｒｅｓｈＳｉｚｅを定義し、指定された代表語判定をおこなう
ためのしきい値を代入する。
部分文書集合の単語頻度用の大きさがｕＷｏｒｄｓの配列変数：ｇｒｏｕｐＦｒｅｑ［ｕＷｏｒｄｓ］を定義する。
繰り返し変数：ｉに１を代入する。
ｗｈｉｌｅ（ｉがＣ以下である）｛
  ｇｒｏｕｐＦｒｅｑのすべての要素に０を代入する。
  繰り返し変数：ｊに１を代入する。
  //　部分文書集合の単語頻度を求める
  ｗｈｉｌｅ（ｊがＮ以下である）｛
    //　部分文書集合に所属する文書かどうかを判定する
    ｉｆ（ＤｏｃＰｏｓ［ｊ］［１］がｉと一致する）｛
      繰り返し変数：ｋに１を代入する。
      ｗｈｉｌｅ（ｋがｄｏｃＰａｒｔｓ［ｊ］［１］のｉＷｏｒｄｓＳｉｚｅ以下である）｛
        ｇｒｏｕｐＦｒｅｑ［ｄｏｃＰａｒｔｓ［ｊ］のｉＷｏｒｄＩＤ［ｋ］］にｄｏｃＰａｒｔｓ［ｊ］の
        ｉＷｏｒｄＦｒｅｑ［ｋ］を加える。
        ｋに１を加える。
      ｝
    ｝
    //　代表語セットの選定
    繰り返し変数：ｊに１を代入する。
    繰り返し変数：ｋに１を代入する。
    ｔｙｐｉｃａｌＷｏｒｄｓＳｉｚｅ［ｉ］に０を代入する。
    ｗｈｉｌｅ（ｊはｕｗｏｒｄｓ以下である）｛
      ｉｆ（ｇｒｏｕｐＦｒｅｑ［ｊ］がｔｈｒｅｓｈＳｉｚｅ以上である）｛
        ｔｙｐｉｃａｌＷｏｒｄｓ［ｉ］［ｋ］を確保する。
        ｔｙｐｉｃａｌＷｏｒｄｓ［ｉ］［ｋ］にｗｏｒｄＮｏｔｅ［ｊ］を代入する。
        ｔｙｐｉｃａｌＷｏｒｄｓＳｉｚｅ［ｉ］に１を加える。
        ｋに１を加える。
      ｝
    ｝
    ｊに１を加える。
  ｝
  ｉに１を加える。
｝
```

【図１９】

文書１

千葉	普通名詞
千葉氏	固有名詞
1950年	普通名詞
フランス	普通名詞
フランス共和国	固有名詞
移住	サ変名詞

文書２

フランス	普通名詞
移住	サ変名詞
希望	サ変名詞
人	普通名詞
増加	サ変名詞

文書３

千葉	普通名詞
商会	普通名詞
チーズ	普通名詞
1950個	普通名詞
発注	サ変名詞

文書４

フランス	普通名詞
フランス産	普通名詞
希望	サ変名詞
小売	普通名詞
価格	普通名詞
1950円	普通名詞

文書５

千葉	普通名詞
千葉産	普通名詞
チーズ	普通名詞
価格	普通名詞
フランス	普通名詞
フランス産	普通名詞
格段	普通名詞
安い	形容詞

文書６

1950年	普通名詞
昼	普通名詞

【図１０】

```
各部分文書集合における関連語セット用の大きさがＣ×可変の２次元配列変数：ｒｅｌａｔｉｖｅＷｏｒｄｓ［Ｃ］
［］を定義する。
各部分文書集合における関連語セット数用の大きさがＣの配列変数：ｒｅｌａｔｉｖｅＷｏｒｄｓＳｉｚｅ［Ｃ］
を定義する。
繰り返し変数：ｉに１を代入する。
総関連語数用変数：ｔｏｔａｌＳｉｚｅを定義する。
ｗｈｉｌｅ（ｉがＣ以下である）｛
  繰り返し変数：ｊに１を代入する。
  ｔｏｔａｌＳｉｚｅに０を代入する。
  ｗｈｉｌｅ（ｊがｔｙｐｉｃａｌＷｏｒｄＳｉｚｅ［ｊ］以下である）｛
    ｉｆ（ｔｙｐｉｃａｌＷｏｒｄ［ｊ］に関連語が登録されている）｛
      関連語辞書からｔｙｐｉｃａｌＷｏｒｄ［ｊ］の関連語数を取得し、ｒｅｌａｔｉｖｅＷｏｒｄｓＳｉｚｅ
    ［ｉ］に代入する。
      繰り返し変数：ｋに１を代入する。
      ｗｈｉｌｅ（ｋがｒｅｌａｔｉｖｅＷｏｒｄｓＳｉｚｅ［ｉ］以下である）｛
        関連語辞書から次の関連語を取得する。
        ／／　同じ表記の単語がすでに関連語として取得されているか否かのチェック
        繰り返し変数：ｌに１を代入する。
        ｗｈｉｌｅ（ｌがｔｏａｌＳｉｚｅ＋ｋ?1以下である）｛
          ｉｆ（取得した単語がｒｅｌａｔｉｖｅＷｏｒｄｓ［ｉ］［ｌ］と一致する）｛
            繰り返しループから抜ける。
          ｝
        ｝
        ｉｆ（ｌがｔｏａｌＳｉｚｅ＋ｋ?1より大きい）｛
          ／／　新しい関連語の登録
          ｒｅｌａｔｉｖｅＷｏｒｄｓ［ｉ］［ｔｏａｌＳｉｚｅ＋ｋ］を確保する。
          ｒｅｌａｔｉｖｅＷｏｒｄｓ［ｉ］［ｔｏａｌＳｉｚｅ＋ｋ］に取得した単語を代入する。
          ｋに１を加える。
        ｝
        ｅｌｓｅ｛
          ／／　すでにある単語は登録しない
          ｒｅｌａｔｉｖｅＷｏｒｄｓＳｉｚｅ［ｉ］から１を引く。
        ｝
      ｝
      ｔｏｔａｌＳｉｚｅにｒｅｌａｔｉｖｅＷｏｒｄｓＳｉｚｅ［ｉ］を加える。
    ｝
    ｊに１を加える。
  ｝
  ｉに１を加える。
｝
```

【図２５】

```
文書データすべてに対し繰り返す｛
  文書データに形態素解析を適用する
  抽出した単語すべてに対し繰り返す｛
    もし抽出した単語がすでに文書データ構成単語リストに存在するならば｛
      文書データ構成単語リストの該当単語の頻度に１を加える
    ｝
    さもなくば｛
      もし抽出した単語がユニーク単語リストに存在しないならば｛
        ユニーク単語リストに該当単語の表記と品詞を登録し、ユニークな識別番号を付与する
      ｝
      文書データ構成単語リストに該当単語の識別番号を追加し、その頻度を１にする
    ｝
  ｝
｝
```

【図１１】

(A)

単語	関連語数	関連語
リナックス	0	
カーネル	0	
OS	0	可能性、可
可能	2	
オペレーティングシステム	0	分配、分与、配給、配布、配達、配当、ディストリビューション
配布	7	日語、ジャパニーズ、国語、母国語
日本語	4	雰囲気、アンビエンス
環境	2	分配、配布、販売、流通
ディストリビューション	4	
Ｖｉｎｅ	0	
商用	0	
アプリケーション	1	アプリ
開発	7	開墾、開発、新開、展開、打開、突破、打破
利用	7	使用、行使、運用、活用、駆使、転用、採用
インストール	1	設置

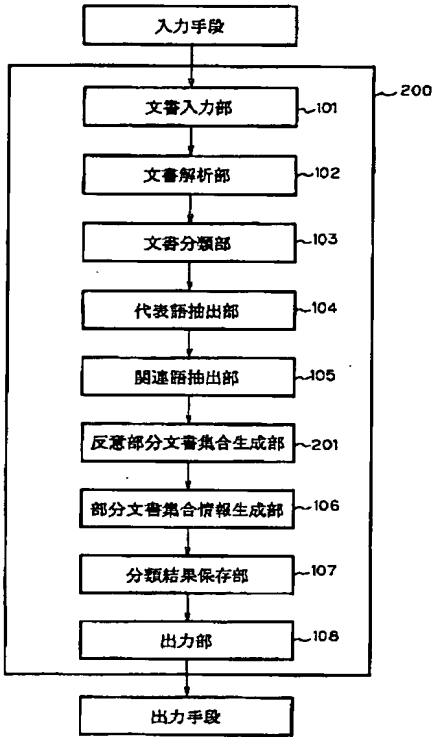
(B)

部分文書集合１	部分文書集合２	部分文書集合３	部分文書集合４
可能性 可 分配 分与 配給 配布 配達 配当 ディストリビューション	日語 ジャパニーズ 国語 母国語 雰囲気 アンビエンス 分配 配布 販売 流通	アプリ 開墾 開発 新開 展開 打開 突破 打破 使用 行使 運用 活用 駆使 転用 採用	設置

【図１２】

```
各部分文書集合における代表語セットはｔｙｐｉｃａｌＷｏｒｄｓ［Ｃ］［］に保持されている。
各部分文書集合における関連語セットはｒｅｌａｔｉｖｅＷｏｒｄｓ［Ｃ］［］に保持されている。
部分文書集合間の代表語セットの和集合用の２次元配列変数：ｔｙｐｉｃａｌＵｎｉｏｎ［Ｃ×（Ｃ?1）／2］［］
部分文書集合間の代表語セットの積集合用の２次元配列変数：ｔｙｐｉｃａｌＩｎｔｅｒｓｅｃｔｉｏｎ［Ｃ×（Ｃ
?1）／2］［］
部分文書集合間の代表語セットの差集合用の３次元配列変数：ｔｙｐｉｃａｌＤｉｆｆ［Ｃ］［Ｃ］［］
部分文書集合間の関連語セットの和集合用の２次元配列変数：ｒｅｌａｔｉｖｅＵｎｉｏｎ［Ｃ×（Ｃ?1）／2］
［］
部分文書集合間の関連語セットの積集合用の２次元配列変数：ｒｅｌａｔｉｖｅＩｎｔｅｒｓｅｃｔｉｏｎ［Ｃ×
（Ｃ?1）／2］［］
部分文書集合間の関連語セットの差集合用の３次元配列変数：ｒｅｌａｔｉｖｅＤｉｆｆ［［Ｃ］［Ｃ］［］

／／　部分文書集合間の代表語セットと関連語セットの和集合、積集合、差集合の生成
繰り返し変数：ｉに１を代入する。
繰り返し変数：ｊに１を代入する。
ｗｈｉｌｅ（ｉがＣ以下である）{
  繰り返し変数：ｋにｉ＋１を代入する。
  ｗｈｉｌｅ（ｋがＣ以下である）{
    ／／　部分文書集合ｉと部分文書集合ｋの代表語セットの和集合の取得
    ｔｙｐｉｃａｌＷｏｒｄｓ［ｉ］とｔｙｐｉｃａｌＷｏｒｄｓ［ｋ］の和集合をｔｙｐｉｃａｌＵｎｉｏｎ［ｊ］
に代入する。
    ／／　部分文書集合ｉと部分文書集合ｋの代表語セットの積集合の取得
    ｔｙｐｉｃａｌＷｏｒｄｓ［ｉ］とｔｙｐｉｃａｌＷｏｒｄｓ［ｋ］の積集合を
    ｔｙｐｉｃａｌＩｎｔｅｒｓｅｃｔｉｏｎ［ｊ］に代入する。
    ／／　部分文書集合ｉと部分文書集合ｋの関連語セットの和集合の取得
    ｒｅｌａｔｉｖｅＷｏｒｄｓ［ｉ］とｒｅｌａｔｉｖｅＷｏｒｄｓ［ｋ］の和集合を
    ｒｅｌａｔｉｖｅＵｎｉｏｎ［ｊ］に代入する。
    ／／　部分文書集合ｉと部分文書集合ｋの関連語セットの積集合の取得
    ｒｅｌａｔｉｖｅＷｏｒｄｓ［ｉ］とｒｅｌａｔｉｖｅＷｏｒｄｓ［ｋ］の積集合を
    ｒｅｌａｔｉｖｅＩｎｔｅｒｓｅｃｔｉｏｎ［ｊ］に代入する。
    ｊに１を加える。
  }
}

繰り返し変数：ｉに１を代入する。
ｗｈｉｌｅ（ｉがＣ以下である）{
  繰り返し変数：ｊに１を代入する。
  ｗｈｉｌｅ（ｊがＣ以下である）{
    ／／　部分文書集合ｉと部分文書集合ｊとの代表語セットの差集合の取得
    ｔｙｐｉｃａｌＷｏｒｄｓ［ｉ］からｔｙｐｉｃａｌＷｏｒｄｓ［ｊ］との共通要素を除いた集合を
    ｔｙｐｉｃａｌＤｉｆｆ［ｉ］［ｊ］に代入する。
    ／／　部分文書集合ｉと部分文書集合ｊとの関連語セットの差集合の取得
    ｒｅｌａｔｉｖｅＷｏｒｄｓ［ｉ］からｒｅｌａｔｉｖｅＷｏｒｄｓ［ｊ］との共通要素を除いた集合を
    ｒｅｌａｔｉｖｅＤｉｆｆ［ｉ］［ｊ］に代入する。
  }
}
```

【図１３】



【図２０】

	文書１	文書２	文書３	文書４	文書５	文書６
千葉	1	0	1	0	1	0
氏	1	0	0	0	0	0
1	1	0	1	1	0	1
9	1	0	1	1	0	1
5	1	0	1	1	0	1
0	1	0	1	1	0	1
年	1	0	0	0	0	1
フランス	1	1	0	1	1	0
共和国	1	0	0	0	0	0
移住	1	1	0	0	0	0
希望	0	1	0	0	0	0
人	0	1	0	0	0	0
増加	0	1	0	0	0	0
商会	0	0	1	0	0	0
チーズ	0	0	1	0	1	0
個	0	0	1	0	0	0
発注	0	0	1	0	0	0
産	0	0	0	1	2	0
小売	0	0	0	1	0	0
価格	0	0	0	1	1	0
円	0	0	0	1	0	0
格段	0	0	0	0	1	0
安い	0	0	0	0	1	0
星	0	0	0	0	0	1

【図１８】

文書１

千葉	普通名詞
氏	固有名詞接尾辞
1	数詞
9	数詞
5	数詞
0	数詞
年	助数詞
フランス	普通名詞
共和国	固有名詞接尾辞
移住	サ変名詞

文書２

フランス	普通名詞
移住	サ変名詞
希望	サ変名詞
人	普通名詞
増加	サ変名詞

文書３

千葉	普通名詞
商会	普通名詞
チーズ	普通名詞
1	数詞
9	数詞
5	数詞
0	数詞
個	助数詞
発注	サ変名詞

文書４

フランス	普通名詞
産	接尾辞
希望	サ変名詞
小売	普通名詞
価格	普通名詞
1	数詞
9	数詞
5	数詞
0	数詞
円	助数詞

文書５

千葉	普通名詞
産	接尾辞
チーズ	普通名詞
価格	普通名詞
フランス	普通名詞
産	接尾辞
格段	普通名詞
安い	形容詞

文書６

1	数詞
9	数詞
5	数詞
0	数詞
年	助数詞
星	普通名詞

【図１４】

```
各部分文書集合における代表語セットはｔｙｐｉｃａｌＷｏｒｄｓ［Ｃ］［］に保持されている。
各部分文書集合における代表語セットの大きさはｔｙｐｉｃａｌＷｏｒｄｓＳｉｚｅ［Ｃ］に保持されている。
各部分文書集合における反対語（関連語）セットはｒｅｌａｔｉｖｅＷｏｒｄｓ［Ｃ］［］に保持されている。
各部分文書集合における反対語セットの大きさはｒｅｌａｔｉｖｅＷｏｒｄｓＳｉｚｅ［Ｃ］に保持されている。

部分文書集合数用変数：ｇｒｏｕｐＳｉｚｅを定義し、Ｃを代入する。
//　ｉは反対語を有する部分文書集合
繰り返し変数：ｉに１を代入する。
ｗｈｉｌｅ（ｉがｇｒｏｕｐＳｉｚｅ以下である）{
  //　ｊは各反対語
  繰り返し変数：ｊに１を代入する。
  ｗｈｉｌｅ（ｊはｒｅｌａｔｉｖｅＷｏｒｄｓＳｉｚｅ［ｉ］以下である）{
    //　ｋは比較対象部分文書集合
    繰り返し変数：ｋに１を代入する。
    ｗｈｉｌｅ（ｋはｇｒｏｕｐＳｉｚｅ以下である）{
      //　ｌは比較体操部分文書集合の代表語
      繰り返し変数：ｌに１を代入する。
      ｗｈｉｌｅ（ｌはｔｙｐｉｃａｌＷｏｒｄｓＳｉｚｅ［ｋ］以下である）{
        ｉｆ（ｒｅｌａｔｉｖｅＷｏｒｄｓ［ｉ］［ｊ］がｔｙｐｉｃａｌＷｏｒｄｓ［ｋ］［ｌ］と一致する）{
          繰り返し変数：ｋのループの外側に抜ける。
        }
        ｌに１を加える。
      }
      ｋに１を加える。
    }
    //　ｋがｇｒｏｕｐＳｉｚｅよりおおきければ、対象の反対語はどの関連語とも一致していない。
    ｉｆ（ｋがｇｒｏｕｐＳｉｚｅより大きい）{
      検索手法を用い、文書群からｒｅｌａｔｉｖｅＷｏｒｄｓ［ｉ］［ｊ］を含む文書を収集する。
      ｇｒｏｕｐＳｉｚｅに１を加える。
      繰り返し変数：ｋに１を代入する。
      ｗｈｉｌｅ（ｋがＮ以下である）{
        ｉｆ（文書ｋが検索された）{
          対象となるＤｏｃＰｏｓ［ｋ］［］に新しい要素を確保し、ｇｒｏｕｐＳｉｚｅを代入する。
        }
      }
      ｔｙｐｉｃａｌＷｏｒｄｓ、ｔｙｐｉｃａｌＷｏｒｄｓＳｉｚｅ、ｒｅｌａｔｉｖｅＷｏｒｄｓ、
      ｒｅｌａｔｉｖｅＷｏｒｄｓＳｉｚｅにそれぞれ要素を一つ追加し、部分文書集合ｇｒｏｕｐＳｉｚｅの代表語セット、関連語セットを求める。
    }
    ｊに１を加える。
  }
  ｉに１を加える。
}
```

【図２６】

```
文書データすべてに対し繰り返す｛
  次元数が文書データ全体でユニークな単語数の文書データベクトルを生成する
  文書データベクトルの全要素に０を代入する
  文書データ構成単語リストのすべての単語に対し繰り返す｛
    保持する識別番号に対応する文書データベクトルの要素に保持する単語頻度を代入する
  ｝
  文書データベクトルを長さ１になるように正規化する
｝
```

【図１７】

文書１

千葉	普通名詞
氏	固有名詞接尾辞
は	副助詞
1	数詞
9	数詞
5	数詞
0	数詞
年	助数詞
に	格助詞
フランス	普通名詞
共和国	固有名詞接尾辞
に	格助詞
移住	サ変名詞
した	サ変助動詞
た	助動詞
。	句点記号

文書２

近年	副詞的名詞
フランス	普通名詞
に	格助詞
移住	サ変名詞
を	格助詞
希望	サ変名詞
する	サ変助動詞
人	普通名詞
が	格助詞
増加	サ変名詞
し	サ変助動詞
ている	接続助詞
。	動詞
	句点記号

文書３

千葉	普通名詞
商会	普通名詞
に	格助詞
チーズ	普通名詞
を	格助詞
1	数詞
9	数詞
5	数詞
0	数詞
個	助数詞
発注	サ変名詞
し	サ変助動詞
た	助動詞
。	句点記号

文書４

フランス	普通名詞
産	接尾辞
の	格助詞
場合	副詞的名詞
、	読点記号
希望	サ変名詞
小売	普通名詞
価格	普通名詞
は	格助詞
1	数詞
9	数詞
5	数詞
0	数詞
円	助数詞
で	格助詞
ある	動詞
。	句点記号

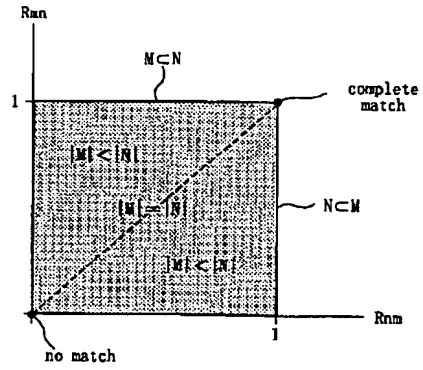
文書５

千葉	普通名詞
産	接尾辞
の	格助詞
チーズ	普通名詞
の	格助詞
価格	普通名詞
は	格助詞
フランス	普通名詞
産	接尾辞
に	格助詞
比べる	動詞
と	格助詞
格段	普通名詞
に	格助詞
安い	形容詞
。	句点記号

文書６

そして	接続詞
、	読点記号
1	数詞
9	数詞
5	数詞
0	数詞
年	助数詞
当時	副詞的名詞
、	読点記号
まだ	副詞
量	普通名詞
を	格助詞
沢山	副詞的名詞
見る	動詞
こと	形式名詞
が	格助詞
出来	動詞
た	助動詞
。	句点記号

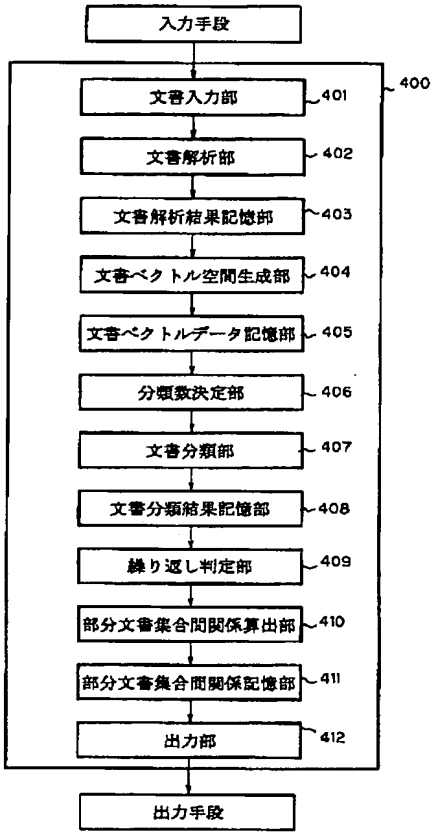
【図２８】



【図２１】

	文書１	文書２	文書３	文書４	文書５	文書６
千葉	1	0	1	0	1	0
千葉氏	1	0	0	0	0	0
1950年	1	0	0	0	0	1
フランス	1	1	0	1	1	0
フランス共和国	1	0	0	0	0	0
移住	0	1	0	0	0	0
希望	0	1	0	1	0	0
人	0	1	0	0	0	0
増加	0	1	0	0	0	0
商会	0	0	1	0	0	0
チーズ	0	0	1	0	1	0
1950個	0	0	1	0	0	0
発注	0	0	1	0	0	0
フランス産	0	0	0	1	1	0
小売	0	0	0	1	0	0
価格	0	0	0	1	1	0
1950円	0	0	0	1	0	0
千葉産	0	0	0	0	1	0
格段	0	0	0	0	1	0
安い	0	0	0	0	1	0
星	0	0	0	0	0	1

【図２４】


入力手段
400
文書入力部
401
文書解析部
402
文書解析結果記憶部
403
文書ベクトル空間生成部
404
文書ベクトルデータ記憶部
405
分類数決定部
406
文書分類部
407
文書分類結果記憶部
408
繰り返し判定部
409
部分文書集合間関係算出部
410
部分文書集合間関係記憶部
411
出力部
412
出力手段

【図２７】

```
分類繰返し数：Ｌを１にする
最大繰返し数をＭにする
繰り返す｛

  文書データベクトルからＬ×Ｎ個のベクトルをランダムに選出し、これらを重心ベクトルとする
  繰り返す｛
    すべての文書データベクトルに対し繰り返す｛
      対象文書データベクトルと最も余弦測度が１に近い重心ベクトルを求める
      上記選出された重心ベクトルを、それと上記対象文書データベクトルとの平均に置き換える
    ｝
    もし、繰返し数が許容値より大きいか、重心と文書データの二乗平均誤差が許容値以下ならば｛
      繰り返しを抜ける
    ｝
  ｝

  すべての重心ベクトルに対し繰り返す｛
    対象重心ベクトルに固有の所属文書データリストを生成する
  ｝

  すべての文書データベクトルに対し繰り返す｛
    対象文書データベクトルと最も余弦測度が１に近い重心ベクトルを求め、
    該当重心ベクトルに固有の所属文書データリストに文書データベクトルの識別番号を加える
  ｝

  Ｌに１を加える
  もし、ＬがＭより大きいならば｛
    繰り返しを抜ける
  ｝
｝
```